

# Signal Processing for Improved Wireless Receiver Performance

**Christensen, Lars P.B.**

*Publication date:*
2007

*Document Version*
Publisher's PDF, also known as Version of record

# Signal Processing for Improved Wireless Receiver Performance

Lars Puggaard Bøgild Christensen

Kongens Lyngby 2007
IMM-PHD-2007-175

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

# Abstract

This thesis is concerned with signal processing for improving the performance of wireless communication receivers for well-established cellular networks such as the GSM/EDGE and WCDMA/HSPA systems. The goal of doing so, is to improve the end-user experience and/or provide a higher system capacity by allowing an increased reuse of network resources.

To achieve this goal, one must first understand the nature of the problem and an introduction is therefore provided. In addition, the concept of graph-based models and approximations for wireless communications is introduced along with various Belief Propagation (BP) methods for detecting the transmitted information, including the Turbo principle.

Having established a framework for the research, various approximate detection schemes are discussed. First, the general form of linear detection is presented and it is argued that this may be preferable in connection with parameter estimation. Next, a realistic framework for interference whitening is presented, allowing flexibility in the selection of whether interference is accounted for via a discrete or a Gaussian distribution. The approximate method of sphere detection and decoding is outlined and various suggestions for improvements are presented. In addition, methods for using generalized BP to perform approximate joint detection and decoding in systems with convolutional codes are outlined. One such method is a natural generalization of the traditional Turbo principle and a generalized Turbo principle can therefore be established.

For realistic wireless communication scenarios, a multitude of parameters are not known and must instead be estimated. A general variational Bayesian EM-algorithm is therefore presented to provide such estimates. It generalizes pre-

viously known methods for communication systems by estimating parameter densities instead of point-estimates and can therefore account for uncertainty in the parameter estimates. Finally, an EM-algorithm for band-Toeplitz covariance estimation is presented as such an estimate is desirable for noise and interference whitening. Using simulations, the method is shown to be near-optimal in the sense that it achieves the unbiased Cramer-Rao lower-bound for medium and large sample-sizes.

# Resumé

Denne afhandling omhandler brugen af signalbehandling til forbedring af trådløse kommunikationsmodtagere i veletablerede cellebaserede netværk som anvendt i bl.a. GSM/EDGE og WCDMA/HSPA. Målet med dette er at forbedre slutbrugerens oplevelse og/eller levere en højere systemkapacitet ved hjælp af øget genbrug af ressourcer.

For at opnå dette mål må man først forstå problemets natur og en introduktion til sådanne systemer er derfor inkluderet. Yderligere gives en introduktion til grafbaserede modeller og approksimationer indenfor trådløs kommunikation sammen med diverse metoder baseret på Belief Propagation (BP), deriblandt Turbo princippet.

Efter at have etableret rammen for forskningen præsenteres diverse metoder til approksimativ detektion. Først introduceres den generelle form for lineær detektion og der argumenteres for at denne form kan være at foretrække f.eks. i forbindelse med parameter estimation. Derefter præsenteres en praktisk metode til hvidtning af støj og interferens, hvilket giver modtageren fleksibilitet i udvælgelsen af om interferens skal modelleres som værende diskret eller Gaussisk fordelt. Den approksimative metode til kugledetektion og -dekodning beskrives og diverse forbedringer til denne foreslås. Herefter introduceres metoder for generaliseret BP i systemer med foldningskoder. En af disse metoder ligger i direkte forlængelse af det traditionelle Turbo princip, hvilket gør det muligt at formulere et generaliseret Turbo princip.

I realistiske trådløse kommunikationssystemer er en masse parametre ukendte og for at estimere disse beskrives en general variationel Bayesiansk EM-algoritme. Denne metode generaliserer hidtil kendte metoder indenfor kommunikationssys-

temer ved at estimere parametrenes sandsynlighedstæthedsfunktion i stedet for det traditionelt anvendte punktestimat, hvilket gør det muligt at tage højde for usikkerhed i parameterestimatet. Endeligt præsenteres en EM-algoritme til estimation af bånd-Toeplitz kovariansmatricer, da et sådant estimat er af interesse til hvidtning af støj og interferens. Det påvises ved hjælp af simuleringer at metoden er nær-optimal for middelstore samt store observationssæt, da den opnår den nedre Cramer-Rao grænse for variansen af centrale estimatorer.

# Preface

The work presented in this thesis was carried out at Informatics and Mathematical Modelling, Technical University of Denmark and at Nokia Denmark A/S in partial fulfillment of the requirements for acquiring the Ph.D. degree in electrical engineering.

The goal of this thesis is to provide a unifying framework of the research carried out in the Ph.D. study during the period Sep. 2003 - Nov. 2006, excluding a leave of abscence from Jan. 2006 - Mar. 2006.

Copenhagen, November 2006

Lars P. B. Christensen

Thesis was successfully defended on the 21/06/2007 with the committee consisting of

**Assoc. Prof. Ole Winther, Technical University of Denmark**

**Prof. Bernard H. Fleury, Aalborg University, Denmark**

**Prof. Hans-Andrea Loeliger, ETH Zürich, Switzerland**

# Contributions

The following publications have been produced during the research study

- [Chr05a] L. P. B. Christensen, *A low-complexity joint synchronization and detection algorithm for single-band DS-CDMA UWB communications*, EURASIP Journal on Applied Signal Processing, UWB - State of the Art, Issue 3, Pages 462-470, 2005.
- [Chr05b] L. P. B. Christensen, *Minimum symbol error rate detection in single-input multiple-output channels with Markov noise*, IEEE SPAWC Workshop, 2005.
- [CL06] L. P. B. Christensen and J. Larsen, *On data and parameter estimation using the variational bayesian EM-algorithm for block-fading frequency-selective MIMO channels*, IEEE ICASSP, 2006.
- [Chr07] L. P. B. Christensen, *An EM-algorithm for band-Toeplitz covariance matrix estimation*, IEEE ICASSP, 2007.

All of the above papers are included with this thesis as appendices. In addition, various more or less novel/useful, but as of yet unpublished, ideas and methods conceived during the research study are outlined below.

- Section 3.4.1: Minimum-phase prefiltered sphere detection and its connection to the QL factorization.
- Section 3.4.2: Cluster sphere detection

- Section 3.5: GBP for improved Turbo equalization in systems with convolutional codes. Based on this, a generalized Turbo principle employing GBP is introduced.
- Section 4.1.3: Exploiting full posteriors for e.g. parameter estimation, not only marginals.

# Acknowledgements

First of all, I would like to thank Nokia Denmark A/S and the Modem System Design group for sponsoring the Ph.D. study. A special thanks goes to Izydor Sokoler and Dr. Søren Sennels for being committed to setting up the Ph.D. study despite challenges to this. I would also like to thank Dr. Niels Mørch for letting me roam around freely in the group, providing me with valuable hands-on experience with real-life algorithms for wireless systems.

During the research study, supervisors involved with the project have been Dr. Thomas Fabricius, Assoc. Prof. Jan Larsen and Dr. Pedro Højen-Sørensen and I would like to thank them all for guiding me through the study and providing valuable input. A special thanks to Pedro for careful reading of this manuscript and many interesting discussions on the topics of this thesis and my sometimes far-fetched ideas. Also thanks to Assoc. Prof. Ole Winther and Prof. Lars K. Hansen for many interesting talks over the years on inference and general signal processing. I would also like to thank the communications and signal processing group at University of California, San Diego for welcoming me during my research visit there. Additionally, I would like to thank Prof. Lars K. Rasmussen for interesting discussions during his research visit at DTU, providing me with a better understanding of loop-correction for GBP.

Finally, I would like to thank my wife Mette for her support and love over the years, in particular when research did not turn out as I had hoped for.

# Ackronyms

| | |
|---|---|
| AWGN | Additive White Gaussian Noise |
| BER | Bit Error Rate |
| BP | Belief Propagation |
| BPSK | Binary Phase Shift Keying |
| CDMA | Code Division Multiple Access |
| CRC | Cyclic Redundancy Check |
| DFE | Decision Feedback Equalization |
| DFT | Discrete Fourier Transform |
| EDGE | Enhanced Data rate for GSM Evolution |
| EM | Expectation Maximization |
| FBA | Forward/Backward Algorithm |
| FDMA | Frequency Division Multiple Access |
| FER | Frame Error Rate |
| FFT | Fast Fourier Transform |
| GBP | Generalized BP |
| GMSK | Gaussian Minimum Shift Keying |
| GSM | Global System for Mobile Communications |
| HSPA | High-Speed Packet Access |
| IIR | Infinite Impulse Response |
| LAN | Local Area Network |
| LDPC | Low-Density Parity-Check |
| LLR | Log-Likelihood Ratio |
| MMSE | Minimum Mean-Square Error |
| LTI | Linear Time-Invariant |
| MAP | Maximum A-Posteriori |
| MIMO | Multiple-Input Multiple-Output |

| | |
|---|---|
| ML | Maximum Likelihood |
| MLSE | Maximum Likelihood Sequence Estimate |
| MMSE | Minimum Mean-Squared Error |
| MSE | Mean-Squared Error |
| PSK | Phase Shift Keying |
| QAM | Quadrature Amplitude Modulation |
| RF | Radio Frequency |
| RRC | Root-Raised Cosine |
| RSSE | Reduced-State Sequence Estimation |
| SNR | Signal-to-Noise Ratio |
| SVD | Singular Value Decomposition |
| TDMA | Time Division Multiple Access |
| VBEM | Variational Bayesian EM |
| WCDMA | Wideband CDMA |
| WP | Weighted Projected |
| ZF | Zero-Forcing |

# Notation

**General Notation**

| | |
|---|---|
| $\mathbf{x}$ | Column vector |
| $x_i$ | Element $i$ of $\mathbf{x}$ |
| $\mathbf{X}$ | Matrix |
| $\mathbf{I}_M$ | Identity matrix of size $M \times M$ |
| $\mathbf{0}_{M \times N}$ | All-zero matrix of size $M \times N$ |
| $[\mathbf{X}]_{i,j}$ | Element $x_{ij}$ of $\mathbf{X}$ |
| $[\mathbf{X}]_{:,j}$ | The $j$'th column of $\mathbf{X}$ |
| $p(\cdot)$ | Probability density of continuous variable |
| $P(\cdot)$ | Probability of discrete variable |
| $\langle f(\cdot) \rangle_{q.(\cdot)}$ | Average of function $f(\cdot)$ over posterior distribution $q.(\cdot)$ |
| $E[\cdot]$ | Ensemble average |
| $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | Complex-valued Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ |
| $\mathcal{CW}^{-1}(\alpha, \boldsymbol{\Sigma})$ | Complex-valued inverse-Wishart distribution with $\alpha$ degrees-of-freedom and covariance $\boldsymbol{\Sigma}$ |
| $\mathcal{X}^2_\alpha$ | Chi-Square distribution with $\alpha$ complex-valued degrees-of-freedom |

**Scalar Operators**

| | |
|---|---|
| $\lvert \cdot \rvert$ | Absolute value |
| $mod(x, y)$ | The value of $x$ taken modulo $y$ |

**Vector Operators**

| | |
|---|---|
| $diag(\cdot)$ | Diagonal matrix given by the vector |

**Matrix Operators**

| | |
|---|---|
| $(\cdot)^*$ | Complex conjugation |
| $(\cdot)^T$ | Matrix transpose |
| $(\cdot)^H$ | Hermitian matrix transpose |
| $\lvert \cdot \rvert$ | Matrix determinant |
| $tr\{\cdot\}$ | Matrix trace, i.e. sum of diagonal elements |
| $\lVert \cdot \rVert$ | Matrix 2-norm |
| $rank(\cdot)$ | Matrix rank |
| $\otimes$ | Kronecker product |
| $diag(\cdot)$ | Vector given by diagonal of the matrix |

**Set Operators**

| | |
|---|---|
| $\mathcal{X} \backslash \mathcal{Y}$ | The set found by removal of $\mathcal{Y}$ from $\mathcal{X}$ |
| $min(\cdot)$ | Minimum of the set |
| $\lvert \cdot \rvert$ | Cardinality of the set |
| $\mathcal{X} \bigcup \mathcal{Y}$ | Union of the sets $\mathcal{X}$ and $\mathcal{Y}$ |
| $\mathcal{X} \bigcap \mathcal{Y}$ | Intersection of the sets $\mathcal{X}$ and $\mathcal{Y}$ |

# Contents

Chapter 1

# Introduction and Motivation

During the last decade, people around the world have embraced wireless communications. Today, nearly everybody in the developed world has a mobile phone or a computer with wireless LAN and the list of potential uses for wireless communications continue to grow. The increasing demand for a better, faster and cheaper wireless experience makes it important for existing systems to be continually optimized in order to improve user experience and remain competitive with upcoming systems. The target of this project is to improve the performance of existing 2G-3.5G cellular systems, such as the GSM/EDGE and WCDMA/HSPA systems deployed throughout Europe and much of the world, within the scope of the already well-established standards and allocated frequency resources.

The performance of a cellular system is a subjective measure, including such quantities as achievable bit-rate, coverage, quality of service and network capacity, all of which depend heavily upon a multitude of variables. However, the scope of this work is only on improvements that can be attributed to the physical layer processing, i.e. processing of signals to and from the antenna sub-system and its effects on objective performance measures such as the Bit Error Rate (BER) or Frame Error Rate (FER). On the other hand, improvements in the physical layer performance will generally improve overall network performance, but in what manner and by how much is a complicated question to answer and is therefore outside the scope of this project.

Besides the goal of providing improved performance, a possible solution must take into account the cost, power and size constraints that an implementation will enforce. This is especially important for the design of a mobile phone as it is highly constrained with respect to both cost, power and size. For example, it may be that a huge gain can be demonstrated by performing optimal processing using multiple antennas, but such a setup is almost guaranteed to be unfeasible due to excessive cost, power and size.

## 1.1 Introduction to Cellular Systems

The application considered in this thesis is cellular communication systems and a quick introduction to the overall functionality of these systems is therefore given here. As implied by the name, a cellular system provides communication services by splitting a given geographical area into cells, each of which has a base-station serving that particular area. The concept of dividing the coverage area into cells is illustrated in Figure 1.1.

However, for communication to take place some amount of resources must be allocated to a particular stream of information. One example of such a resource is the available pool of frequencies in a FDMA network that must somehow be divided between all communication taking place. Conceptually, only one stream of information can occupy a given resource at a time and this will therefore put an upper limit on the total amount of information that can be handled by the system, i.e. its capacity. Fortunately, the idea that resources can only be used once is not the whole truth as there exists a trade-off between the reuse of resources and the achievable bit-rate on a given communication link. Again considering the frequency resources, an example of this strategy for increased capacity can be illustrated by Figure 1.1. The operator of this cellular network could assign a different frequency resource for every cell in a given area and thereby possibly completely eliminating interference between cells or, in the other extreme, use every frequency resource in all cells causing significant interference, but also potentially a major capacity increase. Similar trade-offs exist for all the possible resources available to the network, e.g time-slots in TDMA and codes in CDMA. A different, more traditional strategy for increasing the capacity is to split a cell into smaller cells where required, but this can be costly and is only practical down to a certain cell-size. Operators are therefore naturally interested in being able to tighten the reuse of resources in their network as much as possible and thereby increasing the network capacity or improve the link performance for a given reuse.

To help minimize the interference from the reuse of resources, the network em-

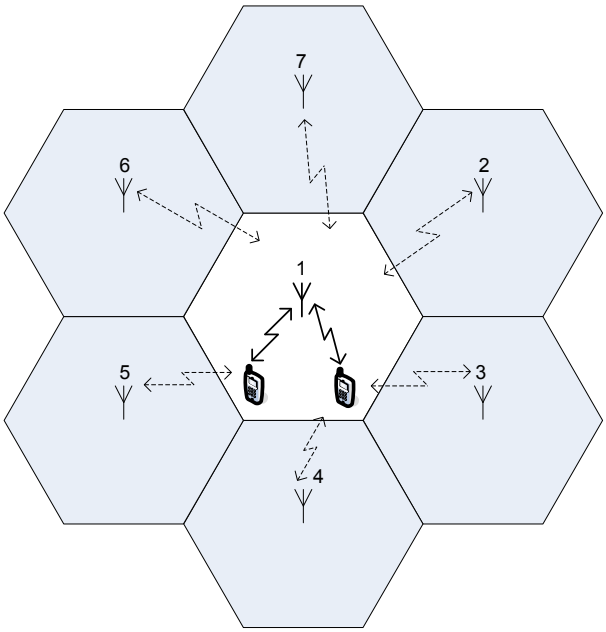


Figure 1.1: Concept of cellular communications.

ploys an adjustable transmission power level known as power-control. Thus when a user is close to a base-station, less power is transmitted to that user and the resulting interference-level to other parts of the system is thereby lowered. Therefore, if a receiver can be designed so that it can handle greater levels of noise and interference, the power-control will simply reduce the power allocated to that stream of information and thereby freeing up the power resource. This then results in higher through-put for users in the network or the possibility of adding more users, i.e. increasing the network capacity.

Besides maximizing capacity, the network should provide as wide coverage as possible in rural areas where the network is typically not capacity-limited, using as few base-stations as possible. For this to be possible, the cell-size should be as large as possible while still maintaining the required bit-rate within a given power budget. In this scenario, the challenge is not so much dealing with interference, but instead extracting the information from a signal with very little power in a noisy environment.

The already mentioned challenges are difficult enough even for an ideal AWGN channel, but radio propagation conditions are typically far from ideal, including significant signal reflections and power fluctuations. Furthermore, users move

around between the cells and support for speeds upwards of 250 km/h are typically required producing significant frequency offsets, i.e. Doppler shifts. The goal of this thesis is therefore to try to meet these challenges and provide possible solutions that can improve the performance of such cellular systems.

## 1.2 Methods of Improving Cellular Performance

Various techniques for improving the physical layer performance of cellular systems can be put into two main categories: Methods requiring changes to the transmitted signal and methods that don't. Examples of the first category are pre-coding of the information in the transmitter and introducing higher-order modulation schemes carrying more bits per symbol. Such methods can be effective, but has the drawback of requiring modifications to the standards and providing backwards compatibility can limit its practical use. The EDGE and HSPA extensions to GSM and WCDMA are examples of this strategy, but this has the drawback of requiring new standards and hardware in order to handle the extensions, all of which adds cost to the network and terminals.

A different strategy, appearing to be getting more focus lately, is that of improving the performance of the receivers employed in the system to allow for a higher degree of resource reuse and possibly better coverage as well. This has the advantage of not requiring any changes to the transmitted signal and can therefore be introduced gradually as networks and terminals are being updated and/or replaced.

However, improving the performance of a receiver under the influence of interference and noise is no easy task. One option is to use multiple receive antennas to effectively provide a better quality signal. Unfortunately, this comes at the price of increased cost, power and possibly size. This may not be a major concern for some applications, but for a mobile phone it can be critical. The focus of this thesis is therefore on improving the receiver performance for a fixed number of antennas, typically one, by improved processing of the observation signal coming from the antenna sub-system.

Chapter 2

# Preliminaries

This chapter builds the basic framework in which the research has been carried out. First, the used system model is presented along with its graph representation. Next, the general topic of inference in graphs is introduced along with its application to the communication system model, including the Turbo principle.

## 2.1 Generic System Model

The wireless communication systems of interest are all of the classical narrowband type operating at a given carrier frequency and the equivalent complex baseband model therefore applies. For a general reference on this topic, see e.g. [Pro95, TV05]. Essentially, this permits the use of traditional linear models for many of the real-life effects on the actual signal.

A schematic overview of the system model is shown in figure 2.1. The transmit structure is split into separate channel encoding and modulation, but more general models with joint encoding and modulation can be constructed to account for various forms of pre-coding, but this is outside the scope of this thesis and has therefore been omitted. In addition, many alternative methods of mapping encoded bits onto a transmitted signal exist, but the linear, memoryless modulation outlined here is either used by the systems of interest or is a good

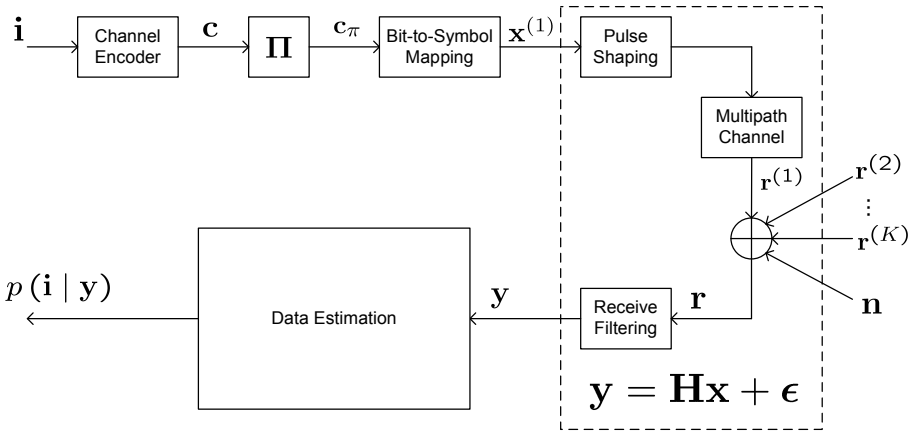


Figure 2.1: Generic wireless system model used throughout the thesis.

approximation and this is therefore the focus of this thesis. In the following it is assumed that user 1 is the only desired user as this is typically the case for a mobile terminal, but the framework can easily be modified to support more than one desired user.

Assuming $N_i$ information bits should be conveyed to the receiver as given by the binary vector $\mathbf{i} \in \{0,1\}^{N_i}$, the task of the channel encoder is to map this information to a new encoded vector $\mathbf{c} \in \{0,1\}^{\frac{N_i}{r}}$ where $0 < r \leq 1$ is the rate of the code. It is often assumed that the input to the encoder is i.i.d. with a uniform distribution, but as the information bits typically come from a source encoder, residual redundancies are likely to be present and thereby violating the assumptions. Additional gains can therefore be achieved by jointly performing the source decoding with the data estimation, but this has the drawback of increased complexity and dependence on the specific type of source and source encoding and this option is therefore not pursued further. The systems of interest typically utilize convolutional codes and it is therefore assumed throughout this thesis that the encoder is a binary convolutional code of rate $r$ having constraint length $N_c$.

Next, the order of the encoded bits are typically permuted by an interleaver to help make the bits appear as independent as possible to the next block, the modulator. Here, bits are collected into blocks of $Q$ bits and mapped onto a complex-valued symbol in the set $\Omega$ out of $|\Omega| = 2^Q$ possible symbols. For example, if $Q = 4$ one could choose to map the bits onto e.g. a 16-QAM or a 16-PSK constellation set. Due to the symbol mapping, the number of transmitted symbols will be $N_x = \frac{N_i}{rQ}$.

Finally, the symbols $\mathbf{x}^{(k)}$ belonging to the $k$'th user are filtered by a pulse-shaping filter to help control the bandwidth of the transmitted signal. A typical choice of pulse-shaping filter is the Root-Raised Cosine (RRC) filter due to its theoretical properties and flexibility, but any filter can in principle be used. The spreading codes used in CDMA systems can be seen as nothing more than a special pulse-shaping filter. This will enforce special properties of the overall pulse-shaping filter that can be exploited, e.g. orthogonality between different codes may be achieved at the expense of excess bandwidth.

The signal is now transmitted across the wireless link by the antenna sub-system. This is accounted for by the time-varying multipath channel that models the effects of reflections and signal fading. However, real-life issues such as timing, frequency-offsets and other RF impairments are not included in this thesis as these effects are typically not a limiting factor in the systems of interest. As discussed previously, interference from other users may occur and the model therefore includes a total of $K$ users. In addition, thermal noise will be present as modeled by the AWGN source $\mathbf{n} \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$.

In the receiver, the signal from the antenna sub-system $\mathbf{r}$ is filtered to produce $\mathbf{y}$ in such a way that all available information about the transmitted bits is preserved in $\mathbf{y}$. Although the text-book answer would be to perform matched-filtering at this point, a real-life implementation depend on the actual system and the environment in which it operates. However, as all operations between the pulse-shaping and the receive filtering are linear operations, the overall transfer function is linear and can be expressed as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon} \tag{2.1}$$

The transfer matrix $\mathbf{H} \in \mathbb{C}^{M\times N}$ is the overall frequency-selective MIMO channel matrix, $\mathbf{x} \in \Omega^N$ is the collection of transmitted symbols from the first $K' \geq 1$ users and $\boldsymbol{\epsilon} \in \mathbb{C}^M$ is the overall noise term containing any remaining users plus filtered AWGN noise. Equation (2.1) looks deceptively simple, but further explanation will follow below in order to better understand it.

Finally, the task of the data estimator in figure 2.1 is to determine the posterior distribution of the information given the observations, as taking decisions based on this distribution will minimize the probability of error [Poo88]. However, for most interesting communication systems, finding this distribution is unfeasible and approximations must be used instead. Such approximations are the topic of this thesis.

Returning to (2.1), the overall channel matrix $\mathbf{H}$ is effectively a linear convolution with temporal dispersion $LT$, where $T$ is the symbol duration. Further, it is assumed that the overall channel coefficients are constant over the considered block of data, i.e. the model is a block-fading model. If the rate of change in the

channel coefficients is so rapid that the block-fading approximation is not valid, this can be accounted for by e.g. a Gaussian state-space model for the channel coefficients [NP03, KFSW02], but this is not considered further in this thesis. For notational convenience, the ramp-up and ramp-down periods of the linear convolution are disregarded as they are typically not of major importance for the overall performance. However, a real-life implementation must take these boundary conditions into account. Based on these assumptions, the resulting structure for the overall channel matrix is

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{L-1} & \cdots & \mathbf{H}_1 & \mathbf{H}_0 & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{L-1} & \ddots & \mathbf{H}_1 & \mathbf{H}_0 & \ddots & \vdots & \vdots \\ \vdots & \mathbf{0} & \ddots & \vdots & \mathbf{H}_1 & \ddots & \mathbf{0} & \vdots \\ \vdots & \vdots & \ddots & \mathbf{H}_{L-1} & \vdots & \ddots & \mathbf{H}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{H}_{L-1} & \dots & \mathbf{H}_1 & \mathbf{H}_0 \end{bmatrix} \tag{2.2}$$

The sub-matrices $\mathbf{H}_l \in \mathbb{C}^{N_r \times K' N_t}$ are the lag $l$ channel matrices with $N_r$ and $N_t$ being respectively the number of receive and transmit dimensions per symbol. Finally, based on the size of the sub-matrices, the size of the overall channel matrix is given by $M = (N_x - L + 1)N_r$ and $N = N_x K' N_t$.

The interference term in the overall noise $\boldsymbol{\epsilon}$ has the same structure as (2.2), only now with an overall channel matrix $\mathbf{H}^{(I)}$ having sub-matrices $\mathbf{H}_l^{(I)} \in \mathbb{C}^{N_r \times (K-K')N_t}$ determining the transfer function from users $\{K'+1, \cdots, K\}$ to the overall noise. The overall noise can therefore be expressed as

$$\boldsymbol{\epsilon} = \mathbf{H}^{(I)} \mathbf{x}^{(I)} + \tilde{\mathbf{n}} \tag{2.3}$$

where $\mathbf{x}^{(I)}$ holds the symbols from users $\{K'+1, \cdots, K\}$ and $\tilde{\mathbf{n}}$ is the thermal noise after receive filtering. Assuming that all transmitted symbols in the overall noise term are i.i.d., zero-mean and unit-power, we have

$$\boldsymbol{\Sigma} \triangleq E\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^H\right] = \mathbf{H}^{(I)} \left(\mathbf{H}^{(I)}\right)^H + \boldsymbol{\Sigma}_{\tilde{\mathbf{n}}} \tag{2.4}$$

where $\boldsymbol{\Sigma}_{\tilde{\mathbf{n}}} \triangleq E\left[\tilde{\mathbf{n}}\tilde{\mathbf{n}}^H\right]$ is the covariance of $\tilde{\mathbf{n}}$ determined by the receive filter. It is then straight-forward to show that $\boldsymbol{\Sigma}$ is a block-banded block-Toeplitz matrix with block-bandwidth $L-1$. The Signal-to-Noise Ratio (SNR) of this system is defined as

$$SNR \triangleq \frac{tr\{\mathbf{H}\mathbf{H}^H\}}{M\sigma^2} \tag{2.5}$$

Under the assumption that $\boldsymbol{\epsilon} \sim \mathcal{CN}(\mathbf{0}_{M\times 1}, \boldsymbol{\Sigma})$, the likelihood of the symbols given the parameters is

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{H}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1} e^{-(\mathbf{y}-\mathbf{H}\mathbf{x})^H \boldsymbol{\Sigma}^{-1} (\mathbf{y}-\mathbf{H}\mathbf{x})} \tag{2.6}$$

For finite systems, the assumption that $\boldsymbol{\epsilon}$ is Gaussian only holds for $K' = K$, but it can serve as a valuable approximation for weak interfering users when $K' < K$. The vast majority of detectors/decoders are most easily derived operating under the influence of AWGN and an equivalent system model fulfilling this requirement is therefore desired. One way of achieving this is to approximate $\boldsymbol{\epsilon}$ as being Gauss-Markov with a memory of $N_m$ symbols, i.e. the block-bandwidth of $\mathbf{\Sigma}^{-1}$ is limited to $N_m$. The closest distribution in the KL-divergence to the original distribution is then found by simply setting elements outside the bandwidth of the inverse to zero [KM00]. By defining the whitening matrix $\mathbf{F}$ by the Cholesky factor $\mathbf{F}^H\mathbf{F} \triangleq \mathbf{\Sigma}^{-1}$ and letting $\tilde{\mathbf{y}} \triangleq \mathbf{F}\mathbf{y}$ and $\tilde{\mathbf{H}} \triangleq \mathbf{F}\mathbf{H}$, we can rewrite (2.6) as

$$p\left(\tilde{\mathbf{y}} \mid \mathbf{x}, \tilde{\mathbf{H}}, \mathbf{F}\right) \propto |\mathbf{F}|^2 e^{-\|\tilde{\mathbf{y}} - \tilde{\mathbf{H}}\mathbf{x}\|^2} \tag{2.7}$$

Again disregarding boundary conditions, the structure of $\tilde{\mathbf{H}}$ is the same as in (2.2), but due to the Gauss-Markov assumption of the overall noise, the effective length of the whitened channel $\tilde{\mathbf{H}}$ is now $\tilde{L} \triangleq L + N_m$ [Chr05b]. This effectively means that any of the considered systems can be transformed into a system of the form

$$\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\mathbf{x} + \tilde{\boldsymbol{\epsilon}} \qquad , \tilde{\boldsymbol{\epsilon}} \sim \mathcal{CN}\left(\mathbf{0}_{M\times 1}, \mathbf{I}_M\right) \tag{2.8}$$

where $\tilde{\boldsymbol{\epsilon}} \triangleq \mathbf{F}\boldsymbol{\epsilon}$. This form of the system model is used throughout the rest of this thesis and a sufficient set of parameters for this system model is then $\boldsymbol{\theta} = \{\mathbf{H}, \mathbf{\Sigma}\}$.

## 2.2 The Channel Capacity and Rate-Diversity Tradeoff

The modern research area of information theory was born with Shannon's ground-breaking theory of communication [Sha48]. Here, the channel capacity is for the first time described as the maximum amount of information carried by a channel such that it can be reliably detected and is found to be

$$C = log_2\left(1 + SNR\right) \qquad [bps/Hz] \tag{2.9}$$

for a scalar channel with AWGN. Designing practical communication systems capable of achieving capacity while having arbitrarily small error probability has been the goal ever since. As realized by Shannon, the channel capacity is easily generalized to multipath channels by frequency-domain water-filling, but it took nearly 50 years before it was generalized to the general MIMO channel

as[1] [Tel99, XZ04]

$$C = N_x^{-1} log_2 \left| \mathbf{I}_M + \sigma^{-2} \mathbf{H} \mathbf{Q} \mathbf{H}^H \right| \qquad [bps/Hz] \tag{2.10}$$

assuming AWGN with $\mathbf{Q} \triangleq E\left[\mathbf{x}\mathbf{x}^H\right]$ determined by water-filling. For fading channels, the so-called ergodic channel capacity can be found by averaging over the distribution of the channel.

For a given fixed channel at high SNR, the ML estimate of the transmitted information has an exponentially vanishing error probability, i.e. $P_e \propto e^{-SNR}$. However considering a fading channel, the probability of error only decays as $P_e \propto SNR^{-d}$, where $d$ is the diversity-order [Pro95, TV05]. For example, if $N$ different observations using independent fading realizations were available, a diversity-order of $N$ could be achieved. There are many ways of achieving this, one possibility being the use of $N$ receive antennas having independent fading between them. Unfortunately, sub-optimal processing may fail to take advantage of the true diversity-order of a system resulting in sub-optimal performance. In general, maximizing the diversity-order is desired to help reliability, but it comes at the price of a reduction in channel capacity compared with that given in (2.10) [ZT03]. Hence, the maximum diversity-order can not be achieved at the rate specified by (2.10) giving rise to a rate-diversity tradeoff. A good example of this is the use of a real-valued modulation such as BPSK on a complex-valued fading channel. To reach capacity, a complex-valued modulation must be used, but the choice of only using half the degrees-of-freedom available results in an increased diversity-order. An example of a similar rate-diversity tradeoff is the choice of using space-time block codes instead of spatial multiplexing for MIMO systems in order to have a higher diversity-order.

As mentioned, sub-optimal processing may fail to extract the available diversity. A good example of this is again the scenario of using BPSK modulation on a complex-valued fading channel. Due to the real-valued modulation, the signal only spans half the signal-space and information should therefore only be extracted from this sub-space. A ML receiver would achieve this whereas the sub-optimal LMMSE detector presented in section 3.1 would not. The reason for this problem is that the complex-valued domain is constrained in the sense that it can only support circular complex-valued distributions, i.e. independent and equal variance real and imaginary components over the complex space. A BPSK modulated signal in a complex-valued channel does not fulfill this circular constraint and the achievable diversity will therefore suffer from the incorrect model. A simple solution to this problem is to map the system onto the unconstrained real-valued domain having twice the number of output dimensions, i.e.

[1] Channel capacity is here defined as the average channel capacity per input symbol over the considered block of data

$$\underbrace{\begin{bmatrix} \mathbf{y}_I \\ \mathbf{y}_Q \end{bmatrix}}_{\mathbf{y}_{IQ}} = \underbrace{\begin{bmatrix} \mathbf{H}_I & -\mathbf{H}_Q \\ \mathbf{H}_Q & \mathbf{H}_I \end{bmatrix}}_{\mathbf{H}_{IQ}} \underbrace{\begin{bmatrix} \mathbf{x}_I \\ \mathbf{x}_Q \end{bmatrix}}_{\mathbf{x}_{IQ}} + \underbrace{\begin{bmatrix} \boldsymbol{\epsilon}_I \\ \boldsymbol{\epsilon}_Q \end{bmatrix}}_{\boldsymbol{\epsilon}_{IQ}} \tag{2.11}$$

where subscript $I$ and $Q$ indicates the real and imaginary part respectively, i.e. $\cdot_I \triangleq Re\{\cdot\}$ and $\cdot_Q \triangleq Im\{\cdot\}$. This representation correctly captures the non-circular statistics of a real-valued modulation and all processing can then be rederived for this modified system. Approximate detectors based on the statistics of the signal can thus extract a greater share of the available diversity in the system [GSL03]. Interestingly, similar structures in space-time block codes can be exploited in the same manner [GOS+04].

## 2.3 Graph Representations and Inference

This section will provide an overview of how the considered system model can be represented and approximated by graphs and thereby help improve the understanding of its underlying structure. The goal of doing this is to exploit the structure of the problem in such a way that inference in these models, e.g. determining hidden variables and parameters, is performed in an efficient manner. This area of research is still very much active and the quest for the ultimate representation of systems as the one depicted in figure 2.1 is still ongoing. The presented graphical framework is based mainly on the work presented in [YFW05], which again builds on decades of research on structured (local) computation. To indicate the versatility of the presented framework, classical methods of increasing generality that can be derived from the framework include: The FFT, forward/backward algorithm, Sum-Product algorithm, Bethe[2] and Kikuchi approximations and the Generalized Distributive Law [AM00]. A related framework is that of Expectation Propagation [WMT05], but this view is not pursued further in this thesis.

### 2.3.1 Factor Graphs and Belief Propagation

A factor graph [KFL01] is a graphical way of expressing how a function of several variables factorizes into functions dependent only on subsets of the variables. For the purpose of this thesis, factor graphs are restricted to representing how

[2] This is the approximation underlying the famed Turbo principle [BGT93, MMC98]

a joint probability distribution function factorizes, i.e.

$$p(\mathbf{x}) \propto \prod_a f_a(\mathbf{x}_a) \tag{2.12}$$

Here $\mathbf{x}_a$ indicates the $a$'th subset of the variables and $f_a$ is a positive and finite function of the subset, so that $p(\mathbf{x})$ is a well-defined distribution. The factor graph contains the structure of (2.12) by a circular variable node for every variable $x_i$ and a square factor node for every function $f_a$. If a given function node $f_a$ depend on $x_i$, an edge will then connect the two. An example of a distribution factorizing in this manner is

$$p(x_1, x_2, x_3, x_4) \propto f_A(x_1, x_2) f_B(x_2, x_3, x_4) f_C(x_4) \tag{2.13}$$

which may be represented by the factor graph shown in figure 2.2. The task

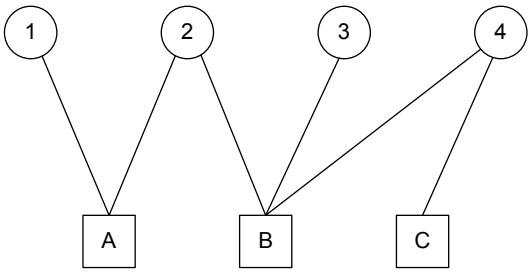


Figure 2.2: Factor graph example.

of computing marginals from distributions of the form given by (2.12) is what we are interested in. For the remaining part of this thesis, it is assumed that all variables in factor graphs are discrete. Although it is possible to have factor graphs with both discrete and continuous variables, e.g. for jointly determining information bits and model parameters, this is outside the scope of this thesis. Letting $S$ be the set of variable nodes that we wish to determine the marginal for, the desired marginal is defined by

$$p_S(\mathbf{x}_S) = \sum_{\mathbf{x} \backslash \mathbf{x}_S} p(\mathbf{x}) \tag{2.14}$$

where the sum over $\mathbf{x} \backslash \mathbf{x}_S$ indicates summing over all combinations of $\mathbf{x}$ not in the set $S$. The problem with performing marginalization as shown in (2.14) is that it requires summing over an exponentially large number of combinations.

The method of Belief Propagation (BP) can help reduce the amount of computations required by exploiting the structure of the problem as represented by the factor graph. However, this may come at the price of marginals only being

approximate, but if the factor graph is loop-free[3], results obtained through BP are guaranteed to converge to their true values once all evidence has been distributed [KFL01]. The graph in figure 2.2 is an example of such a system that has no loops and exact inference can therefore be performed by BP.

The BP algorithm is a message-passing algorithm based on the idea of sending messages from nodes and to its neighbors. The message $m_{a\to i}(x_i)$ from factor node $a$ to variable node $i$ indicates the relative probabilities that $x_i$ is in a given state based on the function $f_a$. Similarly, the message $n_{i\to a}(x_i)$ from variable node $i$ to factor node $a$ indicates the relative probabilities that $x_i$ is in a given state based on the information available to variable node $i$, except for that coming from the function $f_a$ itself. The so-called beliefs, which are simply the approximation to a specific marginal computed by BP, is given by the product of incoming messages and any local factors, i.e.

$$\begin{aligned} b_i(x_i) &\propto \prod_{a\in N(i)} m_{a\to i}(x_i) \\ b_a(\mathbf{x}_a) &\propto f_a(\mathbf{x}_a) \prod_{i\in N(a)} n_{i\to a}(x_i) \end{aligned} \tag{2.15}$$

with $N(i)$ indicating the set of neighbors to node $i$. By requiring consistency using the marginalization condition

$$b_i(x_i) = \sum_{\mathbf{x}_a\setminus x_i} b_a(\mathbf{x}_a) \tag{2.16}$$

the message-updates are found to be

$$\begin{aligned} n_{i\to a}(x_i) &= \prod_{c\in N(i)\setminus a} m_{c\to i}(x_i) \\ m_{a\to i}(x_i) &= \sum_{\mathbf{x}_a\setminus x_i} f_a(\mathbf{x}_a) \prod_{j\in N(a)\setminus i} n_{j\to a}(x_j) \end{aligned} \tag{2.17}$$

The algorithm is sometimes also referred to as the sum-product algorithm due to the lower update of (2.17).

### 2.3.2 Region Graphs and Generalized Belief Propagation

If the factor graph contains loops, the resulting approximation may be far from the exact result, especially if the length of the loop is short. To illustrate this

[3]This means that there is no possible route from any node and back to itself

problem, assume the factor graph in figure 2.2 also has a connection from variable node 3 to factor node C as shown in figure 2.3. There is now a loop[4] in the factor graph as there is a route from variable node 3 and back to itself and BP is therefore not guaranteed to provide exact results. The idea of Generalized Belief Propagation (GBP) is now to propagate messages between regions of nodes instead of single nodes and thereby hopefully providing a better approximation. In figure 2.3, two regions $R_1 = \{A, 1, 2\}$ and $R_2 = \{B, C, 2, 3, 4\}$ have been

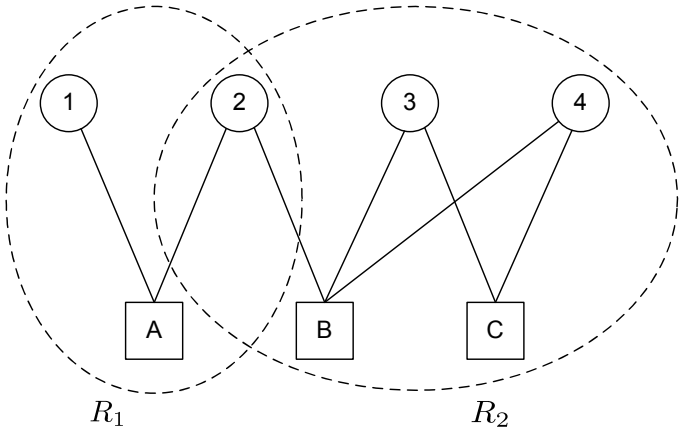


Figure 2.3: Example of region definition on modified factor graph.

defined. Region $R_2$ encapsulates the loop that was causing BP problems and GBP will therefore be exact, but this comes at the price of increased complexity as the complexity scales exponentially with the region sizes. For this little example, the complexity would scale as $\mathcal{O}\left(2^2 + 2^3\right)$ compared with $\mathcal{O}\left(2^4\right)$ for exhaustive search assuming binary variables. However, the real strength of GBP is that even for region definitions that do not encapsulate all loops in the factor graph, the GBP algorithm is still well-defined and can provide improved results compared with BP. Furthermore, through the choice of regions, GBP can scale all the way from BP to exact inference by trading off complexity for improved performance.

In defining the regions, one must ensure that all variables connected to any factor node in the region must also be included in the region. In the example, this results in variable node 2 being included in two regions, but in general nodes may be included in several regions. This raises the question of how communication among regions should be performed, but also the fact that nodes can occur in

[4]It could be argued that this factor graph is in fact loop-free in that merging factor nodes B and C will eliminate the loop without causing a larger complexity. However, this kind of loop encapsulation is not possible for general loopy graphs.

several regions is a concern due to potential over-counting. Region graphs are by definition directed graphs and a possible way to allow communication between regions $R_1$ and $R_2$ is then to define the region $R_3 = R_1 \bigcap R_2 = \{2\}$ and let $R_1$ and $R_2$ be connected to this region. Such a region graph, as shown in figure 2.4, define the interactions between regions and the GBP algorithm operate on such region graphs similarly to how the BP algorithm can be formulated on factor graphs. As was the case for BP on loop-free factor graphs, the GBP algorithm provides exact results when operating on loop-free region graphs [YFW05]. As mentioned before, region $R_2$ encapsulates the loop in the factor graph and the resulting region graph in figure 2.4 is therefore loop-free.

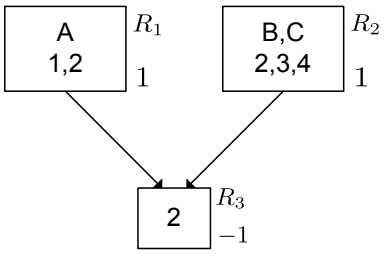


Figure 2.4: Valid region graph for the example.

The potential over-counting of nodes in the factor graph can be dealt with through the use of so-called counting numbers. These counting numbers indicate the weight with which a given region is included in the overall approximation and for the approximation to be well-behaved, the counting numbers of regions involving a given node should sum to one. If $\mathcal{R}$ is the set of all regions each having counting number $c_R$, then the region-based approximation is said to be valid if for all variable nodes $i$ and factor nodes $a$ in the factor graph, we have

$$\sum_{R \in \mathcal{R}} c_R I_R(a) = \sum_{R \in \mathcal{R}} c_R I_R(i) = 1 \tag{2.18}$$

where $I_R(x)$ is a set-indicator function being one if $x \in R$ and zero otherwise. Given the structure of the region graph, it is easy to assign counting numbers that produce a valid approximation. If $A(R)$ is the set of ancestors of a region $R$, then defining the counting numbers as

$$c_R = 1 - \sum_{r \in A(R)} c_r \tag{2.19}$$

will produce a valid region graph. In figure 2.4, the counting numbers associated to each region are also shown and it can be easily verified that the resulting approximation is indeed valid.

Assuming that a given region-based approximation has been specified[5], a GBP algorithm must now be constructed to yield the desired marginals similar to how the sum-product algorithm may be used for regular BP. In fact, there are many such algorithms each generalizing the sum-product algorithm, but here only the so-called parent-to-child algorithm is outlined. The reader is referred to [YFW05] for other possible algorithms.

Advantages of this algorithm are the absence of explicit reference to the counting numbers of the underlying graph and, as the name implies, that it is only necessary to define messages going from parents to their children. In this GBP algorithm, as in regular BP, the belief at any region $R$ can be found by the product of incoming messages and local factors. However, to implicitly correct the potential over-counting, it turns out that we need to include messages into regions that are descendants of $R$ coming from parents that are not descendants of $R$. This is exactly the Markov blanket of region $R$, making the region conditionally independent of any regions other than these. As a result of this, the belief of region $R$ is given by

$$b_R(\mathbf{x}_R) \propto \prod_{a \in A_R} f_a(\mathbf{x}_a) \prod_{P \in \mathcal{P}(R)} m_{P \to R}(\mathbf{x}_R) \prod_{D \in \mathcal{D}(R)} \prod_{P' \in \mathcal{P}(D) \setminus \mathcal{E}(R)} m_{P' \to D}(\mathbf{x}_D) \tag{2.20}$$

where $m_{P \to R}(\mathbf{x}_R)$ is the message from region $P$ to region $R$ and $A_R$ is the set of local factors in region $R$. Furthermore, $\mathcal{P}(R)$ is the set of parent regions to $R$ and $\mathcal{D}(R)$ is the set of descendants with $\mathcal{E}(R) \triangleq R \cup \mathcal{D}(R)$. From (2.20), the message-updates can be found by requiring consistency between parent and child regions yielding

$$m_{P \to R}(\mathbf{x}_R) = \frac{\sum_{\mathbf{x}_P \setminus \mathbf{x}_R} \prod_{a \in A_P \setminus A_R} f_a(\mathbf{x}_a) \prod_{(I,J) \in N(P,R)} m_{I \to J}(\mathbf{x}_J)}{\prod_{(I,J) \in D(P,R)} m_{I \to J}(\mathbf{x}_J)} \tag{2.21}$$

The set $N(P,R)$ consists of the connected pairs of regions $(I,J)$ where $J$ is in $\mathcal{E}(P)$ but not in $\mathcal{E}(R)$ while $I$ is not in $\mathcal{E}(P)$. Further, $D(P,R)$ is the set of all connected pairs of regions $(I,J)$ having $J$ in $\mathcal{E}(R)$ and $I$ in $\mathcal{E}(P)$, but not $\mathcal{E}(R)$.

### 2.3.3 Graph Approximations and Free Energies

Up to this point, it has been assumed that a given graph had somehow been specified as being either an exact or approximate model. First, this section will outline the underlying cost-function that GBP, and hence also BP, minimize. Next, the Bethe and Kikuchi methods of generating approximate graphs are outlined.

[5] How such graphs may be chosen is discussed in the next section

To determine the cost-function of GBP, define the region energy of region $R$ as

$$E_R(\mathbf{x}_R) = -\sum_{a \in A_R} ln[f_a(\mathbf{x}_a)] \tag{2.22}$$

where again $A_R$ is the set of local factors in region $R$. The posterior mean of this energy term is called the region average energy and is naturally given by

$$U_R(b_R) = \sum_{\mathbf{x}_R} b_R(\mathbf{x}_R) E_R(\mathbf{x}_R) \tag{2.23}$$

Also, let the region entropy $H_R(b_R)$ be given by

$$H_R(b_R) = -\sum_{\mathbf{x}_R} b_R(\mathbf{x}_R) ln[b_R(\mathbf{x}_R)] \tag{2.24}$$

allowing us to define the region free-energy $F_R(b_R)$ as

$$F_R(b_R) = U_R(b_R) - H_R(b_R) \tag{2.25}$$

Conceptually, one simply sums up the region free-energies over the entire graph and this is then the metric to minimize. However, due to the over-counting problem, the region free-energies must be weighted by their respective counting number $c_R$ to give the region-based free-energy

$$F_{\mathcal{R}}(\{b_R\}) = \sum_{R \in \mathcal{R}} c_R F_R(b_R) \tag{2.26}$$

where $\mathcal{R}$ is the set of regions in the graph. From (2.26) it can be seen that if the region graph is valid, every variable and factor node from the factor graph is counted exactly once in the region-based free-energy. In [YFW05], the fixed-points of the various GBP algorithms are shown to be fixed-points of the region-based free-energy. What this means is that updating messages according to e.g. (2.21) will locally minimize the region-based free-energy. Furthermore, for the region-based free-energy minimization to make much sense, it must obey some basic constraints. First, the region beliefs $b_R(\mathbf{x}_R)$ must be valid probabilities, i.e. $0 \leq b_R(\mathbf{x}_R) \leq 1$ and sum to one. Additionally, marginals of the region beliefs should be consistent meaning that a marginal should be the same no matter what region it is derived from. If these constraints are fulfilled, the approximation is called a constrained region-based free-energy approximation.

Similar to how the region-based free-energy was found by a weighted sum over the region free-energies, the region-based entropy can be defined in the same way from the region entropies. In [YFW05], it is argued that a good region graph approximation should achieve its maximum region-based entropy for uniform beliefs as the exact region graph must have this property. If a specific region graph fulfills this criteria, it is called a maxent-normal approximation.

### 2.3.4 The Bethe Approximation

An important class of free-energy approximations are those generated by the Bethe method also known simply as Bethe approximations [YFW05]. The region-based approximation generated by this method consists of two types of regions: The set of large regions $\mathcal{R}_L$ and the set of small regions $\mathcal{R}_S$. Any region in $\mathcal{R}_L$ contains exactly one factor node and all variable nodes connected to this factor node. On the other hand, regions in $\mathcal{R}_S$ consists of only a single-variable node and are used to connect large regions having variable intersections. The counting numbers guaranteeing a valid region graph are given by

$$c_R = 1 - \sum_{S \in \mathcal{S}(R)} c_S \tag{2.27}$$

where $\mathcal{S}(R)$ is the set of super-regions of $R$, i.e. regions having $R$ as a subset. Further, all Bethe approximations can be shown to be maxent-normal [YFW05]. Due to the construction of small regions handling the interactions between regions, only single-variable marginals are exchanged and GBP therefore falls back to standard BP on factor graphs. In [YFW05], the Bethe method is generalized to allow multiple factor nodes to be in a region in the large set and similarly regions in the small set are allowed to contain full intersections between regions. This way of generating the region graph is termed the junction graph method and is essentially similar to the generalized distributive law [AM00], which for tree graphs falls back to the famed junction tree algorithm.

### 2.3.5 The Kikuchi Approximation

In the Kikuchi approximation, we use the so-called cluster variation method for generating the regions and associated counting numbers. We start out by a set of large regions $\mathcal{R}_0$ such that every factor and variable node is in at least one region in $\mathcal{R}_0$. Furthermore, no region in $\mathcal{R}_0$ must be a sub-region of another region in $\mathcal{R}_0$. Having defined $\mathcal{R}_0$, the next level of regions $\mathcal{R}_1$ is determined by all possible intersections between regions in $\mathcal{R}_0$, but again making sure that no region in $\mathcal{R}_1$ is a sub-region of another region in $\mathcal{R}_1$. Finally, regions in $\mathcal{R}_0$ are connected to their respective sub-regions in $\mathcal{R}_1$. This process continues until level $K$ where there are no more intersections and the region graph is then given by $\mathcal{R} = \mathcal{R}_0 \cup \mathcal{R}_1 \cup \cdots \mathcal{R}_K$. The counting numbers required to make this a valid region graph is given by (2.27) as for the Bethe approximation.

Unfortunately, region graphs generated by this method are not guaranteed to be maxent-normal. Furthermore, it is argued in [YFW05] that for the free-energy approximation to be good, it should not only be valid and maxent-normal,

but also have counting numbers summing to one when summed over the entire graph. This criteria is not even guaranteed by the Bethe approximation, except for the special case of the graph being loop-free. At present, designing good region-based free-energy approximations that obey even one of these criteria is more of an art than science, but the framework of region-based free-energy approximations is indeed very general and intuitively seems to be a fruitful path for future research. In section 3.5, methods for approximate joint detection and decoding in convolutionally encoded systems is presented based on GBP on region graphs.

### 2.3.6 Helping GBP Converge in Loopy Region Graphs

As for BP, the GBP algorithm is only guaranteed to converge to the exact result when the region graph is loop-free and may even fail to converge for region graphs having multiple loops. A common heuristic for managing this is to let the new message be a convex[6] combination of the update and the last message, either directly on the messages or in the logarithmic domain. There does not appear to be any known theoretical justification for this, but for the systems of interest it seems to work best in the log-domain, i.e.

$$m_{P\to R}^{new}(\mathbf{x}_R) = [m_{P\to R}^{update}(\mathbf{x}_R)]^{w_1}[m_{P\to R}^{old}(\mathbf{x}_R)]^{w_2} \tag{2.28}$$

where $w_2 = 1 - w_1$ and $0 \leq w_1 \leq 1$ is used for convex combining with $w_1$ being a weight factor used to control the update. In fact, this can be seen to be a first-order IIR filter in the log-messages with the IIR filter being provably stable. Obviously, as the weight $w_1$ approaches zero the updates become less important and thereby slowing the convergence of the overall algorithm. On the other hand, doing so stabilizes many, if not all, loopy region graphs as the couplings in the graph are relaxed. Hence, in some sense this scheme seems very similar to that of annealing in that it might be possible to prove that exact inference may be accomplished by letting the convergence rate go to zero and thus effectively perform an exhaustive search [GG84].

An observation that may justify the filtering in log-domain is the over-counting of messages occurring due to loops: If a message $m$ has counted some evidence not once, but $q$ times, the message $m^{1-q}$ should be used instead. This would in fact suggest that the filtering in (2.28) does not necessarily have to be convex, but this raises the question of stability in the log-domain filtering.

Developing a sound theoretical framework for achieving a high probability of convergence for GBP in loopy region graphs while retaining an acceptable com-

[6] A convex combination is a weighting of terms, where all weights are positive and sums to one

plexity remains an open research area. However, the applications of such a framework seem to be numerous as it would be applicable to e.g. general Turbo setups and LDPC codes. In [Yui02], a guaranteed convergent alternative to GBP is presented, but this also comes at the price of much slower convergence. In [LR04, LR05], a filtering scheme operating over the iterations in a Turbo setup is derived assuming that messages are Gaussian and it is shown to provide improved performance. Interestingly, the derived filter is equivalent to the convex IIR filter in (2.28) and based on the Gaussian assumption, an analytical expression for $w_1$ is further provided. Other evidence that such loop-correction schemes may help convergence and hence performance is given in [CC06], where loop-correction is applied to the BP decoding of LDPC codes. Generalizing such ideas to general region graphs and designing methods capable of adaptively compensating for loops while retaining an acceptable complexity seems to be an interesting topic for future research.

## 2.4 Disjoint Detection and Decoding: The Turbo Principle

Determining the exact posterior of the information $p(\mathbf{i} \mid \mathbf{y})$ as shown in figure 2.1 is practically impossible. Even in the unrealistic scenario of known system model and parameters, exact inference would be unfeasible as exhaustive search is the only known method providing exact results in general. This section will therefore describe traditional disjoint detection and decoding performed under the assumption of known system model and parameters and how this can be generalized by the Turbo principle. A common building block of such schemes is the Forward/Backward Algorithm (FBA) used for efficient detection/decoding in systems with memory and this particular algorithm is therefore shortly discussed based on the BP framework introduced earlier.

The idea of disjoint detection and decoding is to separate the two coupled operations by assuming that the other is non-existing and thus resulting in a structure as shown in figure 2.5. First, the input $\tilde{\mathbf{y}}$ is fed into the detector which produces either exactly or approximately the posterior $q(\mathbf{c}_\pi)$ based on the assumption that the coded and interleaved bits $\mathbf{c}_\pi$ are independent apriori. Next, the decoder operates on a deinterleaved version of the posterior called $q(\mathbf{c})$, but in order for the decoder to be simple the input must be independent, i.e. the posterior must factorize as shown in the figure. Therefore, only the marginal posterior is used as this minimizes the KL-divergence[7] under the constraint of full factorization. Based on these marginals, the decoder determines

[7] $D_{KL}(q\|p) \triangleq \left\langle ln\frac{p}{q} \right\rangle_q$

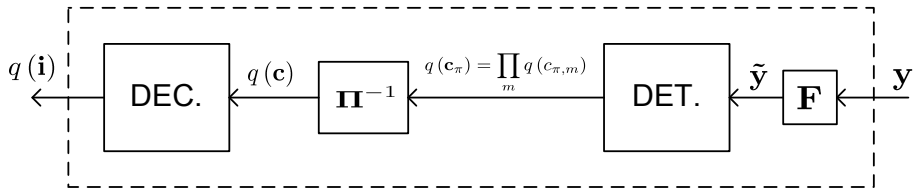


Figure 2.5: Disjoint detection and decoding.

the approximate posterior distribution of the information bits $q(\mathbf{i})$.

The problem with this disjoint scheme is that the detector does not utilize knowledge about the code and the decoder does not use knowledge of the channel. Hence, not all the available structure in the system is taken advantage of and this situation is what the Turbo principle improves upon [BGT93, KST04]. Ideally, the detection and decoding should be performed jointly, but due to complexity constraints this is unfeasible and the basic idea of the Turbo principle is then to iterate between the two disjoint components as illustrated by figure 2.6.

For this to be possible, the detector should be able to accept prior information about the coded bits generated by the decoder. Typically, the decoder can directly produce the desired output and most detectors can be modified to accept priors without too much extra complexity. Instead of propagating the actual bit probabilities between the two components, so-called Log-Likelihood Ratios (LLR) can be more convenient. The LLR $\lambda_i$ of a bit $c_i$ is defined as

$$\lambda_i \triangleq ln\left(\frac{P(c_i = 1 \mid \mathbf{y})}{P(c_i = 0 \mid \mathbf{y})}\right) \tag{2.29}$$

and is therefore just another way of parameterizing the distribution of a bit. To indicate that a fully factorized distribution $q(\mathbf{c}) = \prod_i q(c_i)$ is represented using LLRs, the notation $q_\lambda(\mathbf{c})$ is used. In figure 2.6, it is also shown how the prior input $q_{\lambda,pr}(\mathbf{c})$ to any of the components is subtracted (in LLR domain) from the posterior output $q_{\lambda,p}(\mathbf{c})$ and thereby generating the so-called extrinsic information $q_{\lambda,e}(\mathbf{c})$. This extrinsic information represents the additional information about the coded bits gained by exploiting the structure in the signal at that point, i.e. the channel structure in the detector and the code structure in the decoder. From a graph point of view, the Turbo principle can be seen to be a Bethe approximation [MMC98] and performing BP on this graph will then be equivalent to the structure in figure 2.6. The fact that it is the extrinsic information that should be propagated comes directly from the BP updates in (2.17): Messages going in the opposite direction of a message being updated should not be included in the update. The Turbo framework takes this into account by dividing out the previous information (subtracting the component

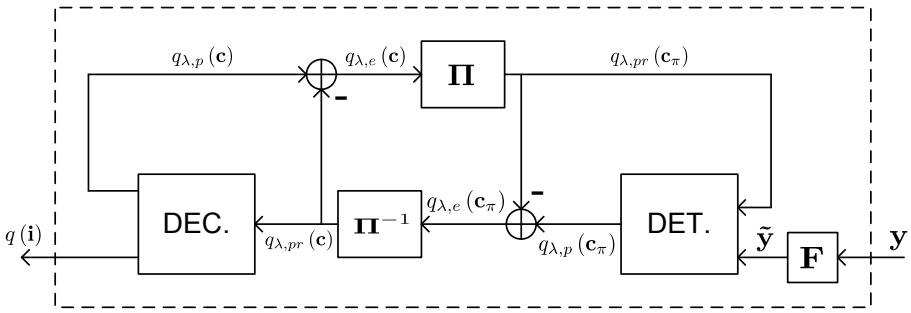


Figure 2.6: Turbo-based detection and decoding.

prior in LLR domain) and thereby forming the component extrinsic information. Due to the uniform prior used by the detector in the first iteration, resulting in $q_{\lambda,pr}(\mathbf{c}_\pi) = \mathbf{0}$, stopping the Turbo iterations after the first decoding will result in the traditional disjoint result. The Turbo principle can therefore be seen as a generalization of the traditional disjoint detection and decoding in figure 2.5.

An important building block for detection and decoding in systems with memory is the FBA [BCJR74], which is simply a special case of the BP algorithm. For disjoint detection and decoding, the algorithm is optimal in the sense that it can determine any desired posterior exactly in an efficient manner by exploiting the Markov structure of a multipath channel or a convolutional code. To illustrate the algorithm, a factor graph for the detection problem can be constructed as shown in figure 2.7, but a factor graph for decoding of convolutional codes will have the same structure. It should be noted that the factor graph is loop-free meaning that using BP on this graph will be exact. In the graph, $x_n$ is a variable sufficient to describe the state of the system at time $n$ and assuming the channel has a temporal length of $\tilde{L}T$, a total of $\tilde{L}$ symbols is therefore required. The most efficient way of distributing the evidence in this graph is by starting at any one point and propagating messages to the ends and back again. This is exactly what the FBA does by defining a forward variable $\alpha_n$ holding information from observations going from left to right, i.e. $\{\tilde{\mathbf{y}}_\mathbf{1}, \cdots, \tilde{\mathbf{y}}_\mathbf{n}\}$ and a backward variable $\beta_n$ going in the opposite direction holding information from observations $\{\tilde{\mathbf{y}}_{\mathbf{n+1}}, \cdots, \tilde{\mathbf{y}}_\mathbf{N}\}$. In the framework of the BP algorithm, the message leaving variable node $n$ and going to the factor node to the right of it would be $\alpha_n$ under the assumption that all nodes to the left of $x_n$ have been updated in a sequential manner. Similarly, the message going in the opposite direction at the same place would be $\beta_n$. Due to the exclusion of messages going in the opposite direction in (2.17), the forward and backward variables will not interact and can therefore be computed separately. The complexity per symbol scales with the set-size of $x_n$ and is therefore $\mathcal{O}\left(N_r 2^{K'\tilde{L}N_t Q}\right)$ where $K'N_t$ is the

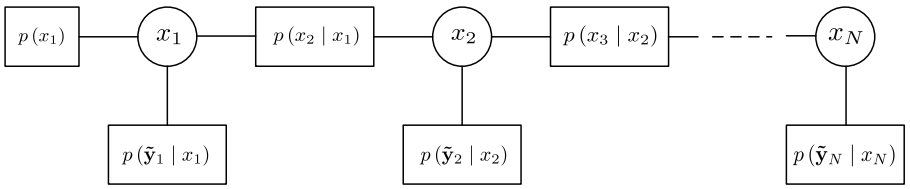


Figure 2.7: Factor graph for Markov model.

effective number of independent users/streams included in the discrete Markov model and $Q$ is the number of bits per symbol. Although this algorithm exploits the Markov structure of the system, the complexity of this detector often makes its implementation unfeasible and approximations must be used instead as will be discussed in chapter 3. For decoding of binary convolutional codes, the complexity per information bit scales as $\mathcal{O}\left(r^{-1}2^{N_c}\right)$ where $r$ is the rate of the code and $N_c$ is the constraint length. Typical real-life values for $r$ and $N_c$ lead to a complexity which is usually implementable and no approximations are therefore required. If the target is not to minimize the information BER but FER, the Viterbi algorithm [Vit67] should be used instead, which again can be seen as a special case of the FBA. Some cellular systems of interest to this thesis use Turbo codes [BGT93] instead of convolutional codes, but these codes are constructed from convolutional codes and the FBA is therefore also used as a component in its decoding. Systems may also employ block codes like LDPC, Reed-Solomon or CRC codes or combinations of all these mentioned coding schemes. Such component codes may also employ the BP algorithm for decoding [MN97, EKM06], but this is outside the scope of this thesis.

Although only described here for iterative exchange of information in a disjoint detector and decoder setup commonly known as Turbo equalization, the Turbo principle is a general way of separating various components from each other such that inference in each component becomes manageable. For example, if the information bits originated from a source encoder, e.g. a voice codec, the Turbo principle could be extended to iterate not only between the detector and the decoder, but also include the source decoder in the iterations.

## 2.5 Summary

This chapter has introduced the signal model that is used throughout this thesis. The model is quite general in the sense that it supports most of the features of interest in today's coded multiuser MIMO systems operating in a multipath environment. For simplicity, the channel model is assumed to be block-fading,

but all methods presented throughout this thesis naturally generalize to a time-varying Gaussian state-space channel model, if desired. Furthermore, the fundamental channel capacity along with the associated rate-diversity tradeoff has been introduced including a IQ-split system model capable of capturing any non-circular properties of the signals.

Next, the concept of representing the system by a factor graph is introduced along with the sum-product algorithm for computing marginals on such graphs. As a generalization of this idea, the region graph is introduced along with a GBP algorithm for computing region marginals and methods for constructing various region graph approximations, namely the Bethe and Kikuchi approximations. In addition, a heuristic method for helping GBP converge in otherwise non-convergent loopy region graphs is presented, but the drawback of this method is slower convergence.

To see how this can be used for disjoint detection and decoding, the now well-known Turbo principle is outlined. The exchange of marginals, called extrinsic information in the Turbo framework, is a direct result of the underlying Bethe approximation leading to a manageable complexity. Having established these methods for detection and decoding, the stage is now set for improving both the individual components of the Turbo scheme, but also going beyond the Turbo framework using more advanced graph approximations.

Chapter 3

# Approximate Detection and Decoding

As optimal detection and decoding is most often infeasible due to the exponentially scaling complexity, this chapter describes various methods of approximate detection and decoding that has been investigated during the research study.

## 3.1 MMSE Detection and Subtractive Extensions

The MMSE detector is, as the name implies, designed to minimize the mean-squared error of the detected signal. The underlying idea of this is, that had the transmitted symbols been Gaussian instead of discrete, the MMSE detector would be optimal. This is a result of the optimality of processing first- and second-order statistics in linear Gaussian models with Gaussian noise, see e.g. [Poo88].

Considering the model in (2.1)[1], the starting point is to assume that the trans-

[1] Noise whitening will be automatically included in the result

mitted symbols are Gaussian and thus specify a prior on the symbols as

$$\mathbf{x} \sim \mathcal{CN}\left(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}\right) \tag{3.1}$$

Using the traditional assumption that the transmitted symbols are i.i.d. with equal probability and unit power, the resulting prior has $\boldsymbol{\mu}_{\boldsymbol{x}} = \mathbf{0}_{N\times 1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathbf{I}_N$. However, if priors are in fact available as in the case of Turbo equalization, the MMSE detector will be able to exploit this extra information [TSK02].

Due to the fact that the Gaussian prior is conjugate[2], the symbol posterior $\hat{\mathbf{x}}$ is also Gaussian, i.e.

$$\hat{\mathbf{x}} \sim \mathcal{CN}\left(\boldsymbol{\mu}_{\hat{\boldsymbol{x}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}}\right) \tag{3.2}$$

with covariance and mean easily found to be given by

$$\begin{aligned}
\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}} &= \left(\mathbf{H}^H \boldsymbol{\Sigma}^{-1} \mathbf{H} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\right)^{-1} \\
\boldsymbol{\mu}_{\hat{\boldsymbol{x}}} &= \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}} \left(\mathbf{H}^H \boldsymbol{\Sigma}^{-1} \mathbf{y} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}}\right) \\
&= \boldsymbol{\Sigma}_x \mathbf{H}^H \left(\mathbf{H} \boldsymbol{\Sigma}_x \mathbf{H}^H + \boldsymbol{\Sigma}\right)^{-1} \left(\mathbf{y} - \mathbf{H} \boldsymbol{\mu}_x\right) + \boldsymbol{\mu}_x
\end{aligned} \tag{3.3}$$

As a result of the Gaussian assumption, the posterior mean $\boldsymbol{\mu}_{\hat{\boldsymbol{x}}}$ is both the MAP and MMSE symbol estimate and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}}$ describe the covariance around this estimate. As in [TSK02], marginals can be generated from this, but (3.3) provides the sufficient statistics of the full posterior distribution and not only the marginals. In fact, the posterior for all detectors should posess a Markov[3] structure as the multipath channel results in $\mathbf{H}^H\mathbf{H}$ being a banded matrix. For this detector, the posterior will not only be Markov, but Gauss-Markov due to (3.2) and $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}}^{-1}$ is therefore a banded matrix.

When used as a component in Turbo equalization, only posterior marginals are required as discussed in section 2.4 and it could therefore be argued that there is no need for the full posterior. However, when incorporating parameter estimation with detection and decoding, the full posterior may in fact be of use as discussed in section 4.1.3.

If desired, other linear and non-linear detectors can be formulated based on this. For example, using $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_M$ and letting $\sigma^2 \to 0$, the Zero-Forcing (ZF) solution is recovered. Another interesting option is to extend the above framework so as to provide subtractive interference cancelation schemes such as the MMSE-DFE with input priors and probabilistic output. Such schemes typically formulate the linear filtering of (3.3) as a forward and a backward filter and then subtract interference terms after some temporal delay, but may also operate across other dimensions, e.g. users in a multiuser scenario. Methods

[2] A conjugate prior results in a posterior of the same form as the prior

[3] Assuming the channel structure is not violated by the noise or prior

such as these are iterative by nature and may actually be viewed as BP on approximate graphs [BC02]. This view is not explicitly considered further in this thesis, but prefiltered sphere detection as described in section 3.3 can be considered a generalization of this idea with interference being subtracted once a sufficient level of confidence in a symbol has been reached.

## 3.2 Detection with Whitening

The basic idea of modeling weak interferers as Gaussian noise was introduced in connection with the system model in section 2.1 and is as such a straightforward method of interference rejection. However, more details are required to understand how practical implementations of this scheme can be constructed.

In the generic system model, the whitening filter $\mathbf{F}$ was found by Cholesky factorization of the inverse noise covariance $\mathbf{\Sigma}^{-1}$ limited to have a bandwidth of $N_m$ symbols. However, even for known symbols $\mathbf{x}$, estimating $\mathbf{\Sigma}$ from the signal $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{H}\mathbf{x}$ is no easy task as the sample covariance matrix $\boldsymbol{\epsilon}\boldsymbol{\epsilon}^H$ is rank-one. This problem can be overcome by enforcing a banded Toeplitz structure to the estimate as the true covariance matrix is known to be of this form [Chr07]. However, the complexity of such inherently iterative schemes may be too great and alternative methods operating on a smaller, better conditioned covariance matrix is of interest.

A way of achieving this is to process only a sliding-window of the received signal at a time and let the Forward/Backward Algorithm (FBA) handle interactions between the overlapping sections [Chr05b]. Let $\boldsymbol{\epsilon}_{n-N_m}^{n}$ be the sliding-window at time $n$ from the noise signal $\boldsymbol{\epsilon}$. The length of the take-out window is $N_m + 1$ symbols and the window moves in steps of one symbol time $T$. As $N_m$ is typically small compared to the number of observations, the covariance of $\boldsymbol{\epsilon}_{n-N_m}^{n}$ can be reliably estimated as

$$\mathbf{\Sigma}_{N_m} \triangleq E\left[\boldsymbol{\epsilon}_{n-N_m}^{n}\left(\boldsymbol{\epsilon}_{n-N_m}^{n}\right)^H\right] \simeq \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\epsilon}_{n-N_m}^{n}\left(\boldsymbol{\epsilon}_{n-N_m}^{n}\right)^H \tag{3.4}$$

Furthermore, due to the properties of the sliding-window, the resulting covariance estimate will be well-behaved having a Toeplitz structure (see e.g. [Chr07] and references therein). In [Chr05b], the FBA is derived for optimal detection for such a system. The main result of this is that a special whitening filter can be derived from $\mathbf{\Sigma}_{N_m}$ such that the ordinary FBA without overlap in the observations can be applied, but due to the memory of the noise the effective channel length grows from $L$ to $L + N_m$. However, by employing the smaller

covariance matrix of (3.4) in the covariance estimation, the band-constraint is effectively imposed on the covariance matrix itself and not its inverse. Hence, even assuming perfect covariance estimates, this scheme is not equivalent to filtering with the true $\mathbf{F}$, but as $N_m$ becomes "large" compared to the true bandwidth of the covariance matrix, it will approach the result obtained from filtering with $\mathbf{F}$. This is a result of $\mathbf{\Sigma}_{N_m}^{-1}$ approaching the diagonally centered sub-matrices of $\mathbf{\Sigma}^{-1}$ as $\mathbf{\Sigma}_{N_m}$ captures more and more of the structure in $\mathbf{\Sigma}$. The

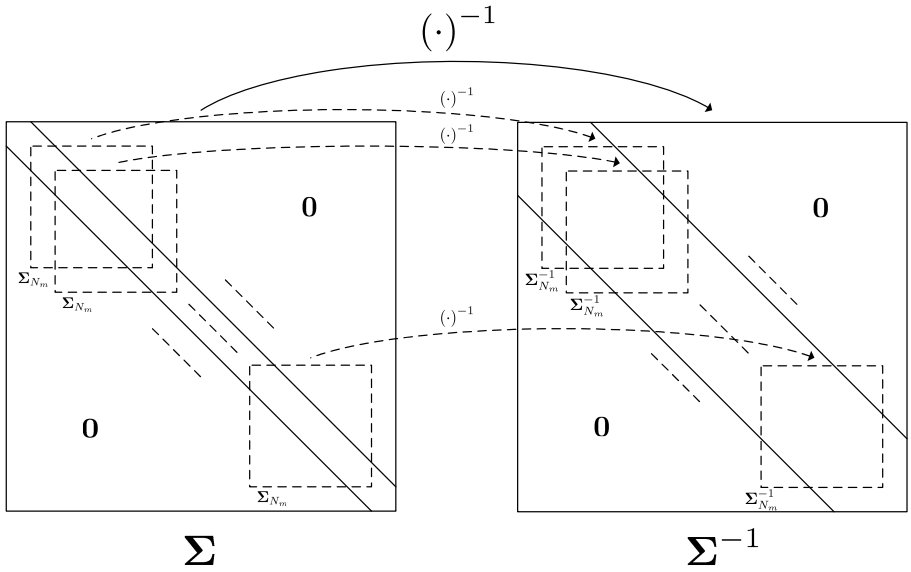


Figure 3.1: Relation between $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{N_m}$ and their respective inverses.

relation between the two covariances $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{N_m}$ and their respective inverses is illustrated by figure 3.1 with fully drawn lines indicating $\mathbf{\Sigma}$ and dotted lines $\mathbf{\Sigma}_{N_m}$. As previously mentioned in the section on the generic system model, $\mathbf{\Sigma}^{-1}$ is only strictly Toeplitz when disregarding the boundary conditions, i.e. for infinite systems, and the same limitation also applies to the "sliding" of the inverse covariance $\mathbf{\Sigma}_{N_m}^{-1}$ in the right-hand side of figure 3.1.

To illustrate the potential of the described whitening solution in the presence of interference, simulations of a GSM link is shown in figure 3.2. The Carrier-to-Interference Ratio (CIR) is defined as the ratio between the desired signal power $C$ and the total interference power $I = \sum_i I_i$. The left-hand side of the figure shows the BER for a single interferer ($K = 2$) and the right-hand side shows performance for two interferers ($K = 3$) having a 10dB power difference. The used channel model is the so-called TUx multipath model [3GP] defined for GSM and resulting in an overall channel of length $L = 7$. The simulations are

done using perfect parameter knowledge, i.e. perfect knowledge of channel and covariance. Although classical sampling theory suggests that the oversampling factor $N_{sps}$ does not have to be higher than one [Pro95], it may be advantageous to use oversampling in practice due to various challenges such as synchronization, excess bandwidth in transmit pulseshaping, adjacent channel interference, etc. The difference in performance in figure 3.2 going from $N_{sps} = 1$ to $N_{sps} = 2$ is due to the additional channel diversity that may be exploited from the excess bandwidth of the transmitted signals. Another point that should be stressed here is the IQ-split of the covariance matrix as introduced in section 2.2. As

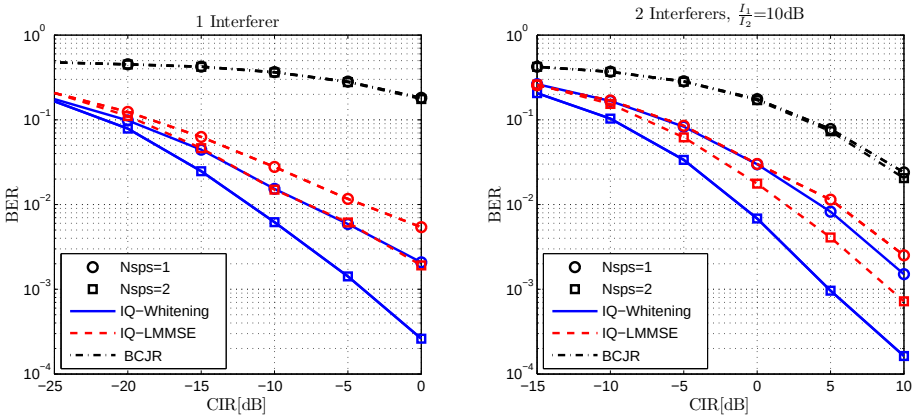


Figure 3.2: BER of a GSM link in the presence of interference, $N_m = 3$, $K' = 1$, SNR=40dB.

mentioned earlier, this method is capable of greatly increasing the achievable diversity for real-valued modulations such as the GMSK[4] modulation used in GSM.

However, the important point here is the difference in performance between different detectors for the same setup. The FBA derived for an AWGN channel, also known simply as the BCJR algorithm due to the authors of [BCJR74], is indicated by BCJR in the figure. Due to the AWGN assumption of the detector, performance lacks greatly compared to the two detectors incorporating knowledge of the noise covariance, i.e. MMSE and whitening. It should also be noted how whitening consistently outperforms the MMSE detector, but the drawback of whitening is naturally its potentially much larger complexity due to using the FBA on a channel of length $\tilde{L} = L + N_m$. If the increased complexity is allowed, whitening can potentially outperform any other detector having an equal or smaller set of discrete signals in its model, e.g. the MMSE, as whitening performs optimal inference in the discrete set of signals. Another advantage

[4] The GMSK modulation used in GSM may be well-approximated by a $\frac{\pi}{2}$-BPSK modulation

of this is, that it does so without the need for additional parameter estimates compared with the MMSE, e.g. channel coefficients of interfering signals. This is important in a real-life implementation as reliably estimating channel coefficients of all interfering users is often virtually impossible. The real strength of whitening is therefore the capability of scaling the set of discrete signals in the model all the way from the Gaussian MMSE and to full discrete joint detection. A natural extension of this is then to further approximate the detection in the discrete set by the use of suboptimal detectors and thus provide an even more flexible detection framework. A possible solution for approximate detection in the discrete set is discussed next.

## 3.3 Sphere Detection and Decoding

An intriguing method of reduced-complexity approximate inference is by sphere detection and decoding (see e.g. [HV05a, HV05b, DEGC03, AEVZ02] and references therein). The basic idea of this method is to consider all possible transmitted signals as points in a multi-dimensional space and then simply search for the point closest to the received signal point with the added constraint that searching is only carried out inside a sphere around the received signal point. This has the advantage of potentially eliminating a lot of candidates from the otherwise exhaustive search, but also the risk of not finding any points inside the sphere if the radius is too small.

To illustrate the principle of sphere detection and decoding, figure 3.3 shows the detection problem for a frequency-flat real-valued 2x2 MIMO channel using BPSK modulation. It can be seen how the lattice of possible transmission points on the left becomes skewed by transmission through the channel matrix $\tilde{\mathbf{H}}$. The task is now to determine the point closest to the received point at the center of the red circle and thus determining the MLSE symbol solution. This also illustrates the difficulty of choosing a good radius for the search: If the radius is reduced much compared to the one in the figure, no points will be found whereas increasing the radius will include more points in the sphere and thus having a higher complexity.

Once a given radius of the sphere has been specified, the points inside the sphere can be determined by a bounding technique relying on a triangular representation of the channel matrix by either a QR or QL factorization. Let $\tilde{\mathbf{H}} = \tilde{\mathbf{Q}}\tilde{\mathbf{R}}$ where $\tilde{\mathbf{Q}}$ is unitary and $\tilde{\mathbf{R}}$ upper triangular. Due to the AWGN assumption, the metric to minimize for MLSE is

$$C(\hat{\mathbf{x}}) = \|\tilde{\mathbf{y}} - \tilde{\mathbf{H}}\hat{\mathbf{x}}\|^2 = \|\dot{\mathbf{y}} - \tilde{\mathbf{R}}\hat{\mathbf{x}}\|^2 \leq r^2 \tag{3.5}$$

with $\dot{\mathbf{y}} \triangleq \tilde{\mathbf{Q}}^H \tilde{\mathbf{y}}$ and $r$ being the radius of the sphere. This modified system is effectively determined by the upper triangular matrix $\tilde{\mathbf{R}}$ making it possible to bound one dimension at a time. Taking the 2x2 MIMO system as an example, we have

$$C(\hat{\mathbf{x}}) = \left\| \dot{\mathbf{y}} - \tilde{\mathbf{R}}\hat{\mathbf{x}} \right\|^2 = \left\| \begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} - \begin{bmatrix} \tilde{r}_{11} & \tilde{r}_{12} \\ 0 & \tilde{r}_{22} \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} \right\|^2 \\ = |\dot{y}_1 - \tilde{r}_{11}\hat{x}_1 - \tilde{r}_{12}\hat{x}_2|^2 + |\dot{y}_2 - \tilde{r}_{22}\hat{x}_2|^2 \tag{3.6}$$

Using $C(\hat{\mathbf{x}}) \leq r^2$, we readily determine the bound for $\hat{x}_2$ as

$$\frac{\dot{y}_2 - r}{\tilde{r}_{22}} \leq \hat{x}_2 \leq \frac{\dot{y}_2 + r}{\tilde{r}_{22}} \tag{3.7}$$

Next, fixing the first dimension, which in this case is $\hat{x}_2$, allows us to set

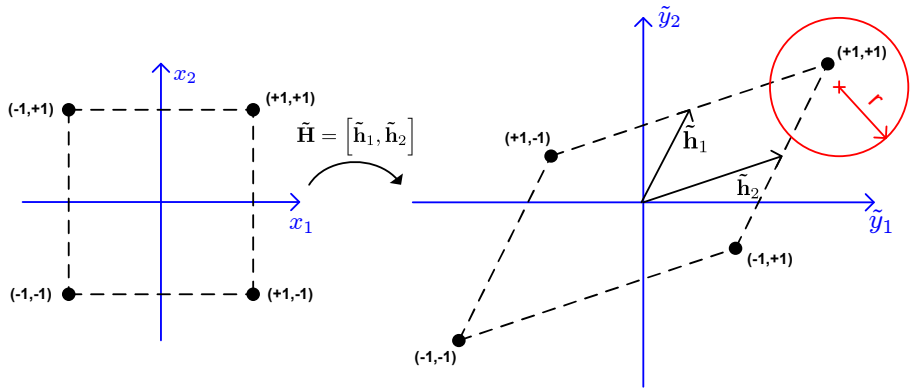


Figure 3.3: Example of sphere detection for a simple real-valued 2x2 MIMO system with BPSK modulation.

up bounds for the next in the same way and so on until all dimensions have been included. Although only directly applicable to real-valued modulations[5], the outlined bounding technique has a straight-forward generalization to the complex-valued domain involving circle intersections instead of a circle intersecting a line as in (3.7). Assuming $\tilde{\mathbf{H}} \in \mathbb{C}^{M \times N}$ and letting $p \triangleq rank\left(\tilde{\mathbf{H}}\right)$, the bounding technique described can only provide bounds for $p$ dimensions and $p \geq N$ must therefore apply if all symbol dimensions should be bounded. In the case of $p < N$, bounds can only be provided for $p$ dimensions and exhaustive search in the remaining $N - p$ dimensional subspace is therefore required to guarantee the MLSE solution.

[5]This includes complex modulations that may be decomposed as a superposition of orthogonal real modulations, e.g. rectangular QAM modulations

Turning the attention towards predicting the radius of the sphere, we rewrite (3.5) using $\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\mathbf{x} + \tilde{\boldsymbol{\epsilon}}$ as

$$C(\hat{\mathbf{x}}) = \|\tilde{\mathbf{Q}}^H \tilde{\boldsymbol{\epsilon}} + \tilde{\mathbf{R}}(\mathbf{x} - \hat{\mathbf{x}})\|^2 \simeq \|\tilde{\boldsymbol{\epsilon}}\|^2 \tag{3.8}$$

with the approximation being good when the MLSE solution is close to the transmitted signal, i.e. at high SNR. As $\tilde{\boldsymbol{\epsilon}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$, the squared norm is $\|\tilde{\boldsymbol{\epsilon}}\|^2 \sim \mathcal{X}^2_M$ where $\mathcal{X}^2_M$ indicates a Chi-Square distribution with $M$ complex-valued degrees-of-freedom. Based on this, one could then choose to select the radius such that there is a certain probability $\varepsilon$ that the transmitted point is not in the sphere, i.e.

$$P\left(\|\tilde{\boldsymbol{\epsilon}}\|^2 > r^2\right) = \varepsilon \tag{3.9}$$

Due to the Chi-Square distribution of $\|\tilde{\boldsymbol{\epsilon}}\|^2$, this can be easily accomplished by look-up tables for the distribution. For high SNR regimes, this method works very well and huge reductions compared to exhaustive search can be achieved, but as the SNR decreases so does the pruning of the search space. This is a result of the sphere constraint becoming less strict as a relatively large radius is needed to provide a high probability of having any points inside the sphere.

An advantage of the sphere detection and decoding framework described so far is, that if any points are found inside the sphere, the MLSE solution is guaranteed to be in this set of points. However, a major drawback is that the bounds determined as in (3.7) tend to be very loose in the first dimensions. A natural solution to this problem is to generalize the single radius constraint to a set of increasing radii [GH03], but this has the drawback of not guaranteeing the MLSE solution. However, a great reduction in complexity is achievable by this and the resulting BER can come arbitrarily close to MLSE performance by accepting a higher complexity.

Interestingly, the multipath model of (2.2) makes sphere detection possible without performing a full QR or QL factorization. The easiest way to illustrate this is by including the ramp-up and -down parts of the channel matrix in $\tilde{\mathbf{H}}$, now making it block lower triangular instead of upper. To match the structure of $\tilde{\mathbf{H}}$, a QL factorization is therefore considered instead of a QR factorization. If the block size is one, no QL factorization is required at all, but in general we have

$$\tilde{\mathbf{H}} = \left(\tilde{\mathbf{Q}}_0 \otimes \mathbf{I}\right) \tilde{\mathbf{L}} \tag{3.10}$$

where $\tilde{\mathbf{H}}_0 = \tilde{\mathbf{Q}}_0 \tilde{\mathbf{L}}_0$ is the lag 0 channel matrix of the multipath channel. This effectively makes it possible to apply the sphere constraint of (3.7) on top of the Viterbi algorithm and thereby guaranteeing the MLSE solution by searching only inside the sphere [VHM06].

Until this point, only the search for the MLSE solution has been discussed in connection with sphere detection and decoding, but in a coded system a soft-input

soft-output detector is usually the goal and not the MLSE solution. However, the desired bit posterior may be approximated from the evaluated points inside the sphere [HtB03, VHK04], but if there are too few points evaluated the approximation can naturally be poor.

## 3.4 Improved Sphere Detection

This section presents two ideas that may help the understanding and performance of current state-of-the-art sphere detectors known to the author. The ideas should not be thought of as final results that are ready for publication, but more as potentially fruitful future research paths.

### 3.4.1 Minimum-phase Prefiltering and the QL Factorization

A classical result of Linear Time-Invariant (LTI) systems is the spectral factorization theorem stating that any spectrum can be factorized into minimum-phase components [SK01]. The finite-length equivalent of this is given by the Cholesky factorization of any covariance matrix [ADC95]. Further, the spectral factorization for LTI systems has an extension stating that any filter may be factorized into an all-pass filter and the minimum-phase filter given by spectral factorization. This is a crucial result in connection with Reduced-State Sequence Estimation (RSSE), as it can be shown that a minimum-phase filter has the highest possible energy concentration in the first taps [EQ88, DHH89]. A common trick is therefore to filter the incoming signal with the conjugate of the all-pass filter associated with the channel, resulting in a minimum-phase channel. As the energy is concentrated in the first taps, the minimum-phase characteristic gives the best possible starting point for taking early decisions. However, the generalization of this result to finite-length systems appear to be unknown[6] and this section will try to argue that such an extension exists as given by the QL factorization of the channel matrix.

The channel matrix of interest here has a structure as given in (2.2), including the ramp-up and -down parts if desired. To simplify things, a block-size of 1 is assumed to recover the classical LTI system, but all arguments have natural block-level extensions allowing for larger block-sizes. The system size is assumed $\mathbf{H} \in \mathbb{C}^{M \times N}$ with a channel length of $L$ symbols and $M = N + L - 1$. The QL

[6] At the time of writing, the author has not been able to find any references on this.

factorization of this channel is uniquely given by

$$\mathbf{H} = \mathbf{QL} \tag{3.11}$$

where $\mathbf{Q} \in \mathbb{C}^{M \times M}$ is unitary and $\mathbf{L} \in \mathbb{C}^{M \times N}$ is the lower triangular positive definite Cholesky factor, i.e.

$$\mathbf{H}^H \mathbf{H} = \mathbf{L}^H \mathbf{L} \tag{3.12}$$

Due to boundary conditions, the matrix $\mathbf{H}^H \mathbf{H}$ is not guaranteed to be Toeplitz, but the finite channel length limits its to bandwidth $L$ and the same is therefore also true for $\mathbf{L}$. Letting $N \to \infty$ while keeping $L$ fixed, it can be shown that the columns (or rows) of $\mathbf{L}$ must be shifted versions of each other and is exactly given by the spectral factorization [ADC95]. An intuitive argument for this is that any given column interacts in the same way with its neighboring columns as the boundary effects have vanished. Turning the attention to the all-pass filter, the matrix equivalent of this must be a unitary matrix as this is the only "filter" leaving the spectrum/covariance unchanged. As $\mathbf{Q}$ is the only unitary matrix relating $\mathbf{H}$ and $\mathbf{L}$, this must represent the all-pass filter associated with the spectral factorization. Again, for $N \to \infty$ the argument of vanishing boundary effects can be invoked and all columns of $\mathbf{Q}$ must therefore be shifted versions of each other and tend to the all-pass filter. For finite systems, boundary conditions will influence the result and Toeplitz structures are therefore not guaranteed. However, for large enough systems, results will approach their asymptotic values in the middle of the system where boundary conditions are less significant.

This apparent connection between minimum-phase prefiltering and the QL factorization of convolutive channels makes it possible to directly link the area of minimum-phase prefiltered RSSE to that of sphere detection. Through the connection with the QL factorization, a theoretically sound method of exploiting minimum-phase prefiltering in sphere detection can be constructed generating tighter bounds. Hence, sphere detection on multipath channels can be seen to be very similar to that of RSSE with the difference being that decisions are not made until a given degree of certainty has been achieved as determined by the radii used. An intriguing idea is then to construct a unifying framework of the two worlds by combining the pruned state-space search of sphere detection with the decision-feedback world of RSSE and related methods, see e.g. [BC02]. The author believes this to be a fruitful topic of future research.

### 3.4.2 Cluster Sphere Detection

At present, the sphere detection framework appear to be an attractive candidate for high to medium SNR regimes, but may fail to provide the desired

performance for low to medium SNR scenarios when limited to a realistic complexity. A major reason for this failure can be seen from the Singular Value Decomposition (SVD) of the channel matrix

$$\tilde{\mathbf{H}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^H \tag{3.13}$$

where $\tilde{\mathbf{U}} = [\tilde{\mathbf{u}}_1, \cdots, \tilde{\mathbf{u}}_M] \in \mathbb{C}^{M\times M}$ and $\tilde{\mathbf{V}} = [\tilde{\mathbf{v}}_1, \cdots, \tilde{\mathbf{v}}_N] \in \mathbb{C}^{N\times N}$ are unitary matrices holding the left and right singular vectors of the channel and the diagonal matrix $\tilde{\mathbf{\Sigma}} \in \mathbb{R}^{M\times N}$ contains the singular values $[\tilde{\sigma}_1, \cdots, \tilde{\sigma}_m]$, where $m \triangleq min(\{M, N\})$. The singular values are non-negative and in decreasing order, i.e. $[\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_m]$. Multiplying symbols onto the channel matrix can therefore be expressed as

$$\tilde{\mathbf{H}}\mathbf{x} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^H\mathbf{x} = \sum_{i=1}^{m} \left(\tilde{\sigma}_i \tilde{\mathbf{v}}_i^H \mathbf{x}\right) \tilde{\mathbf{u}}_i \tag{3.14}$$

where each term of the sum in (3.14) is a subspace component of the overall result. The strength of each of these subspace components is determined by the projection of $\mathbf{x}$ onto $\tilde{\mathbf{v}}_i$ scaled by $\tilde{\sigma}_i$. Hence, if $\tilde{\sigma}_i^2$ is small compared to the noise variance, which in this model is unity, little information can be carried by this subspace or sub-channel [Tel99]. Let signal points $\mathbf{x}_x$ and $\mathbf{x}_y$ have difference vector $\mathbf{d}_{x,y} \triangleq \mathbf{x}_x - \mathbf{x}_y$. Assuming that $\mathbf{d}_{x,y}$ has a large portion of its energy in directions with small singular values, the signal points will then be clustered close to each other relative to the noise standard deviation.

Figure 3.4 illustrates this problem by letting vectors $\tilde{\mathbf{h}}_1$ and $\tilde{\mathbf{h}}_2$ approach each other and thereby reducing $\tilde{\sigma}_2$. Selecting the radius based on the noise statistics is therefore likely to give a sphere that includes both points close to the center as they are "close" relative to the selected radius. As both points are almost equally likely to have produced the observation, this artifact is therefore not a failure in the radius selection, which correctly predicts the set of points that must be evaluated to get an accurate posterior. However, for multi-dimensional systems with higher-order modulations, these clustering effects are problematic due to the near-exhaustive search that inherently must be carried out inside such a cluster of candidate points in order to produce an accurate posterior.

A possible solution to this problem is to approximate clusters of points as being Gaussian instead of consisting of a number of discrete points. Such an approximation would have several appealing properties, the first being that if no Gaussian approximation is made, the sphere detection framework described so far is recovered. Furthermore, for $SNR \to \infty$ no reasonable system exists that will experience the clustering effect and traditional sphere detection is therefore also recovered here. For $SNR \to 0$, only a single Gaussian cluster exists resulting in the Gaussian approximation of (3.3), which is known to be asymptotically

optimal [PV97]. In between these two extremes, the approximation is a hybrid of discrete and Gaussian components depending on the criteria used for defining the clusters.

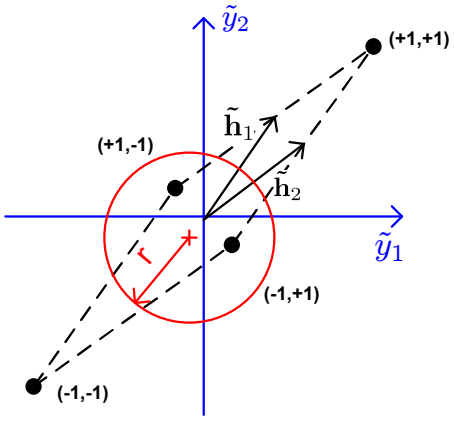


Figure 3.4: Illustration of sphere detection for a poorly conditioned channel with low to medium SNR.

Let $\mathcal{X}_n$ denote the set of signal points included in the $n$'th Gaussian cluster given by its mean $\boldsymbol{\mu}_n \in \mathbb{C}^N$ and covariance $\boldsymbol{\Sigma}_n \in \mathbb{C}^{N\times N}$. Assuming that $N_{cl}$ clusters exist, then $\mathcal{X} \triangleq \bigcup_{n=1}^{N_{cl}} \mathcal{X}_n$ is the collection of all signal points included through Gaussian clusters. The posterior of this approximate system will then have terms coming from the discrete signal points and terms from the Gaussian components, i.e. the posterior is of the form

$$\begin{aligned} p\left(\hat{\mathbf{x}} \mid \tilde{\mathbf{y}}, \boldsymbol{\theta}\right) &\propto p\left(\tilde{\mathbf{y}} \mid \hat{\mathbf{x}}, \boldsymbol{\theta}\right) p\left(\hat{\mathbf{x}}\right) \\ &= p\left(\tilde{\mathbf{y}} \mid \hat{\mathbf{x}}, \boldsymbol{\theta}\right) \left( \sum_{\hat{\mathbf{x}}_d \in \Omega^N \setminus \mathcal{X}} P\left(\hat{\mathbf{x}}_d\right) \delta\left(\hat{\mathbf{x}} - \hat{\mathbf{x}}_d\right) + \sum_{n=1}^{N_{cl}} Z_n p\left(\hat{\mathbf{x}} \mid \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right) \right) \end{aligned} \tag{3.15}$$

where $\delta\left(\cdot\right)$ is the Kronecker delta function. The left-most sum in (3.15) is the discrete contribution to the posterior and this part should be well-approximated by the already described sphere detection framework by searching over the discrete space $\Omega^N \setminus \mathcal{X}$. The terms in the right-most sum of (3.15) are the Gaussian components given by

$$p\left(\hat{\mathbf{x}} \mid \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right) = \left|\pi \boldsymbol{\Sigma}_n\right|^{-1} e^{-\left(\hat{\mathbf{x}} - \boldsymbol{\mu}_n\right)^H \boldsymbol{\Sigma}_n^{-1} \left(\hat{\mathbf{x}} - \boldsymbol{\mu}_n\right)} \tag{3.16}$$

The normalization constant $Z_n$ in (3.15) scales each individual Gaussian com-

ponent so that integrating over it recovers the associated discrete prior, i.e.

$$Z_n = \sum_{\hat{\mathbf{x}}_d \in \mathcal{X}_n} P\left(\hat{\mathbf{x}}_d\right) \tag{3.17}$$

This makes sure that no matter the approximation used, the overall hybrid prior is proper and integrates to one.

The approximation of (3.15) assumes knowledge of the Gaussian clusters and determining these while keeping the complexity low is an open problem of this method. Clearly, simply evaluating all points is not an option as this would be equivalent to exhaustive search. Returning to the difference vector $\mathbf{d}_{x,y}$, we define the difference set $\mathcal{D}$ such that $\mathbf{d}_{x,y} \in \mathcal{D}$ if

$$\|\tilde{\mathbf{H}}\mathbf{d}_{x,y}\|^2 = \|\tilde{\mathbf{R}}\mathbf{d}_{x,y}\|^2 \leq r_{\mathcal{D}}^2 \tag{3.18}$$

Here, $r_{\mathcal{D}}^2$ is the maximum squared distance allowed inside a single cluster and should be chosen based on the noise statistics so that there is "little" impact of the approximation, e.g. $r_{\mathcal{D}}^2 \simeq E\left[\|\tilde{\boldsymbol{\epsilon}}\|^2\right] = M$. The condition of (3.18) is a sphere detection problem and the set $\mathcal{D}$ can therefore be determined efficiently by this method. However, performing the clustering based on $\mathcal{D}$ and determining the mean and covariance of each of these clusters in an efficient manner is still an open problem of this approximate method.

## 3.5 Approximate Joint Detection and Decoding using GBP

Until now, this chapter has dealt only with detectors that can be employed in a Turbo-based receiver and not considered any other approximations to the problem of joint detection and decoding. However, based on the region-based free-energy approximations described in section 2.3.2, this section will present methods for performing approximate joint detection and decoding of convolutional coded signals over multipath channels, i.e. the system model shown in figure 2.1. As the well-known concept of Turbo equalization for such a problem is equivalent to the Bethe approximation, the basic idea is to use a more advanced graph approximation to hopefully provide better performance without incurring the exponential complexity of an exhaustive search. A similar approach was taken in [PA06] for joint detection and decoding, but instead of convolutional codes, LDPC coding was considered.

### 3.5.1 The Modified Cluster Variation Method

The basic concept of this method is based on the cluster variation method for constructing a Kikuchi approximation as described in section 2.3.5. An example of a region graph generated by this method for the system of interest is shown in figure 3.5. Here, the top-level regions in $\mathcal{R}_0$ each contain one observation and

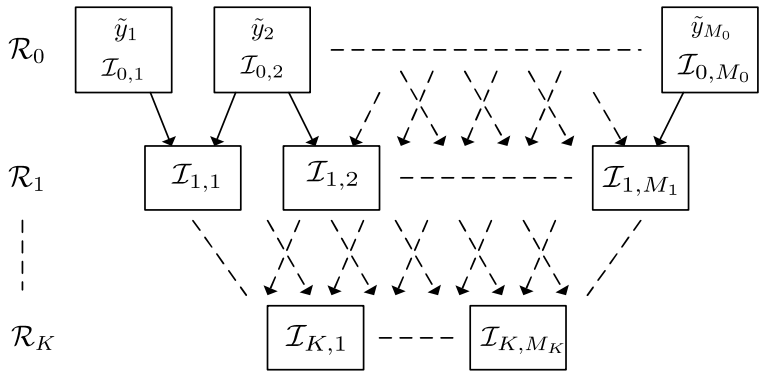


Figure 3.5: Region graph found by the cluster variation method for convolutionally coded signal over multipath channel.

the corresponding set of information bits required for conditional independence of the observation, e.g. $\tilde{y}_1$ is given by the set of information bits $\mathcal{I}_{0,1}$. Naturally, the sets of information bits associated with each of the observations depend on the rate of the code, the interleaver and the length of the channel, but given these parameters the graph is deterministic and can be predetermined. The next levels of regions are then found by the cluster variation method until no regions intersect in the final $K$'th layer. The counting numbers for regions in $\mathcal{R}_0$ are all set to one and the remaining counting numbers are found by (2.27), guaranteeing a valid region graph. Due to the Markov structure in both the convolutional code and the channel, the set-size of regions in the top-level is no larger than $N_cLQ$, i.e. $|\mathcal{I}_{0,j}| \leq N_cLQ$ for all $j$, where $N_c$ is the constraint length of the code, $L$ is the length of the channel in symbols and $Q$ is the number of bits per symbol. Furthermore, as a result of the cluster variation method using intersections to form regions, we have

$$|\mathcal{I}_{i,j}| \leq N_cLQ - i \qquad ,\forall j \tag{3.19}$$

The maximum number of levels in the region graph is therefore $N_cLQ$, i.e. $K \leq N_cLQ - 1$. Only counting the top-level regions, this method will therefore have a complexity in the order of $\mathcal{O}(r^{-1}2^{N_cLQ})$ per information bit per iteration whereas the complexity of a Turbo equalization iteration would be

$\mathcal{O}(r^{-1}2^{N_c} + 2^{LQ})$. Due to the many connections between layers, the region graph will generally be far from loop-free and the computed beliefs are therefore only approximations. However, by merging of top-level regions this approximate joint detection and decoding scheme can scale all the way to exact inference, but as the complexity of the GBP algorithm scales exponentially in the region size, doing so results in a greatly increased complexity culminating in an exhaustive search when all top-level regions are merged into one.

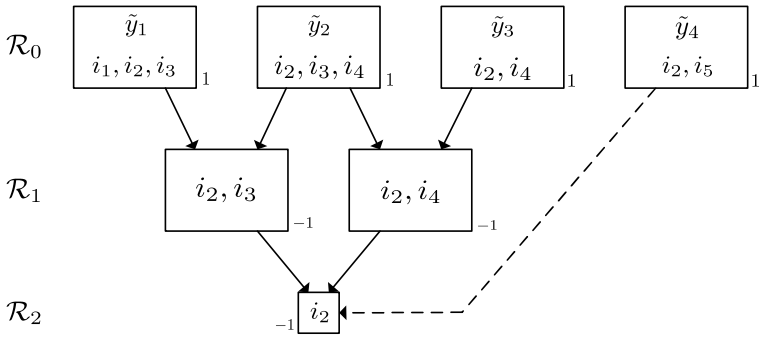


Figure 3.6: Example of a loopy region graph generated by the modified cluster variation method.

As mentioned in section 2.3.3, the beliefs should be constrained in the sense that marginals should be consistent no matter what region they are derived from. For this to be possible, the region graph must allow regions containing the same variables to communicate their intersection, i.e. they must have at least an indirect connection with the intersection being a subset. Let's consider the example in figure 3.6 where the regions in $\mathcal{R}_1$ and $\mathcal{R}_2$ are found according to the cluster variation method. However, the right-most top-level region will not be connected at all for this method as its intersection with the other top-level regions is a sub-region of a region in $\mathcal{R}_1$ and the marginal of $i_2$ will therefore not be consistent. The cluster variation method is therefore modified to tackle this problem by connecting any unconnected variables to regions at lower levels, or create such required regions as necessary, so that communication can take place. We will call this the modified cluster variation method and for the example, using this modified method will result in the creation of the dotted connection. It seems obvious that the unconnected region should be connected in this simple example, but in the general case one should realize that any unconnected variable should be connected to other regions involving the same variable, either directly or indirectly. When using this modified cluster variation method, the counting numbers are found as usual and in the example, counting numbers associated with each region are shown next to that region.

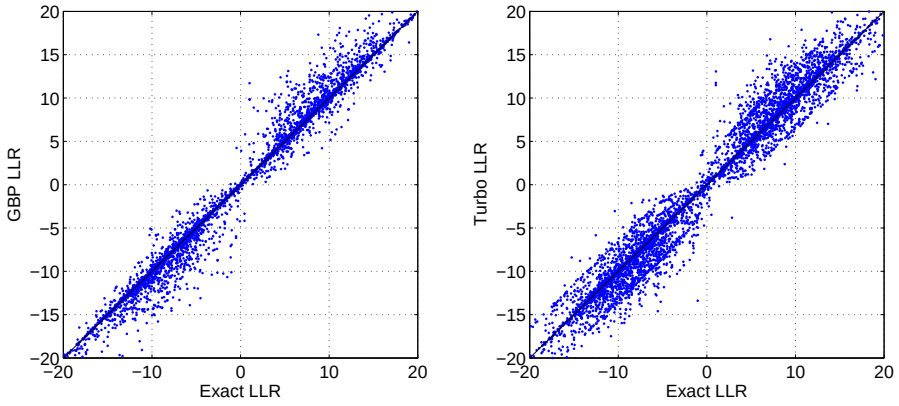


Figure 3.7: Comparison of LLRs for the modified cluster variation method and Turbo equalization, $SNR = 3dB$, $w_1 = 0.1$

An important observation from figure 3.6 is that the region graph has a loop and that the resulting marginals are therefore not exact. For this simple example, it is possible to simply merge the regions in $\mathcal{R}_1$ resulting in a loop-free region graph, but this is not a general solution as this results in an exponentially increasing complexity. Generally, region graphs constructed by this method will have many loops as can be seen from figure 3.5. Unlike the graph approximation underlying the Turbo principle, the resulting loop length from using this method does not increase with system size as this will simply produce a wider graph in figure 3.5. Hence, increasing the system size will not make it more probable that GBP converges and heuristic loop-correction as proposed in section 2.3.6 must instead be relied upon to make the GBP algorithm convergent. Unfortunately, this results in slower convergence, which is only practical up to a certain point and it therefore seems that this method is not a generally viable solution.

However, to show that the method does in fact work when employing sufficient loop-correction, a simple system using a rate $r = \frac{1}{2}$ convolutional code with generator polynomial $g(D) = [1, 1 + D]$ is considered resulting in a constraint length of $N_c = 2$. The system transmits blocks consisting of $N_i = 12$ information bits and random interleaving is employed. This is then mapped onto BPSK symbols and transmitted over a multipath channel of length $L = 2$ given by $\mathbf{h} = [1, \frac{1}{\sqrt{2}}]^T$. For this system, loop-correction using the convex IIR filtering in (2.28) with $w_1 = 0.1$ seems to provide convergence with probability one. In figure 3.7, the LLRs computed by the GBP algorithm on the region graph found by the modified cluster variation method is plotted against the exact value found by exhaustive search for 500 blocks. Also shown for comparison is the result found by traditional Turbo equalization for the same realizations of interleavers

and observations and it can be seen that GBP generally improves the quality of the marginals compared to that achieved by Turbo equalization. Although not shown here, the LLRs derived from GBP generally also result in a BER closer to that of the exact result, but marginals being closer to the exact result is in itself a desired quality, e.g. for parameter estimation as discussed in chapter 4.

It would be interesting to try out this method for larger systems having higher constraint and channel lengths, but as the complexity scales as $\mathcal{O}(r^{-1}2^{N_c LQ})$ this is only feasible for small values of $N_c LQ$. However, the perhaps biggest obstacle to such an approach seems to be the slow convergence coming from the loop-correction required to guarantee convergence. However, using the cluster variational method to form the region graph appear to yield good results for some special detection problems [SWS04].

### 3.5.2 The Generalized Turbo Principle

Inspired by the failure of the modified cluster variation method due to an excessive amount of short loops in the region graph, a natural way of avoiding such loops is to take a closer look at the Turbo principle. Focusing on Turbo equalization, the structure of the underlying graph approximation can be illustrated as shown in figure 3.8. Here, the lower and upper Markov chains represent respectively the channel and convolutional code, with the crossing connections representing the interleaved exchange of extrinsic information. An important property of this structure is that for random interleaving, the probability of short loops decreases with the system size, i.e. the interleaver length [XES01]. This is a general property of the Turbo principle and is one of the main reasons why this framework has been so successful. However, for smaller sized systems the over-counting resulting from loop-feedback can result in inferior performance and non-convergence and it is for such systems that we will try to improve upon the Turbo principle.

In fact, under the assumption of random interleaving, the graph structure asymptotically approaches a tree [AV01, XES01] and the Turbo principle is therefore asymptotically optimal. For finite systems, loops of finite length will exist and it is for such systems that we will try to improve upon the Turbo principle.

Unlike the modified cluster variation method where the notion of Markov chains is lost, we will now explicitly preserve the two Markov chains, i.e. as given by the channel and convolutional code. In fact, the structure in figure 3.8 can be viewed as two Markov chains which intersect each other in a manner determined by the interleaver. The Turbo principle then lets these two Markov

chains exchange single-variable beliefs in the form of extrinsic information at the intersection points. The general idea is now to modify the exchange of information so that not only single-variable beliefs are exchanged, but entire region beliefs. This idea of exchanging region beliefs between components, in this example between Markov chains, readily generalizes to any scheme where the Turbo principle can be employed and will therefore be called the generalized Turbo principle. Random interleaving in such a system will scale the loop-length with the system size in the same manner as for the ordinary Turbo principle and thus providing the same desirable features leading to a high probability of convergence. In the framework of region graphs, the graph approximation underlying such an approach can be seen to be a junction graph as the exchange of information is no longer accomplished by single-variable beliefs, but by multi-variable beliefs [YFW05]. In essence, one thereby exchanges not single-variable beliefs (or scalar-valued extrinsic information) between components, but multi-variable beliefs (or vector-valued extrinsic information) over the defined regions.

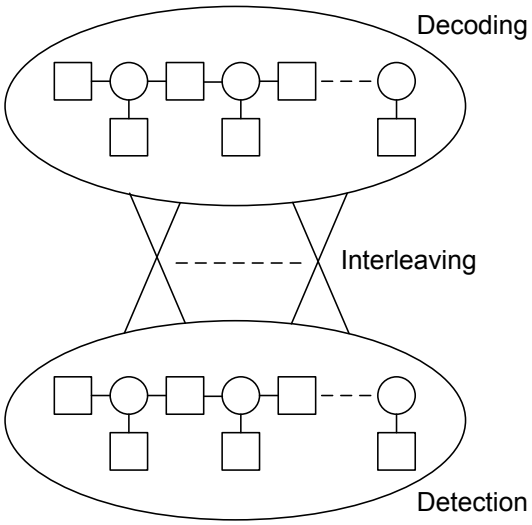


Figure 3.8: Illustration of the underlying graph structure of Turbo equalization.

For the Turbo equalization system considered, such a method should be able to capture more of the dependency between the two Markov chains. To accomplish this, regions should ideally capture the full intersection between the Markov chains as represented by a maximum of $N_c LQ$ information bits leading to a complexity of $\mathcal{O}(r^{-1}2^{N_c LQ})$ instead of $\mathcal{O}(r^{-1}2^{N_c} + 2^{LQ})$ for ordinary Turbo equalization. However, an interesting option is to only capture the strongest couplings between the Markov chains and thus establishing a framework in which

one can tradeoff performance for lower complexity. In such a framework, constraining the exchange of beliefs to being single-variable beliefs recovers the traditional Turbo principle.

Setting up the region graph starts out as in the cluster variation method, i.e. defining regions in $\mathcal{R}_0$ to be the observations and associated information bits. Here it should be noted that as in the cluster variation method, the number of information bits required in each region in $\mathcal{R}_0$ is a maximum of $N_c LQ$ bits. The next level $\mathcal{R}_1$ will be defined so as to handle the channel interactions as can be accomplished by choosing regions to be the intersections between any two neighboring regions in $\mathcal{R}_0$, i.e. in the time-domain Markov chain. Similarly, re-

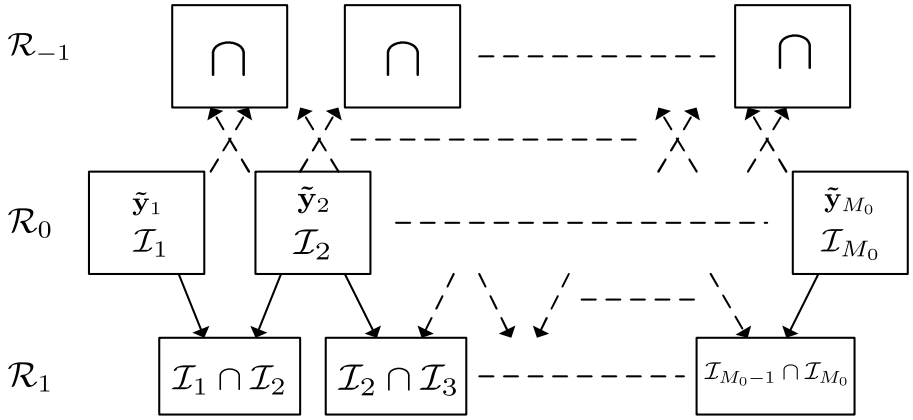


Figure 3.9: Region graph structure for generalized Turbo equalization.

gions in $\mathcal{R}_{-1}$ represent the convolutional code Markov chain and are also found as intersections between any two neighboring regions in $\mathcal{R}_0$. However, neighboring regions should here be seen from the convolutional code Markov chain point-of-view and determining which regions are neighbors of a region is therefore uniquely given by the code rate and the interleaver. An illustration of this method of constructing the region graph is shown figure 3.9. Here, the definition of regions in $\mathcal{R}_{-1}$ is not explicitly shown due to regions being determined by intersections of regions in $\mathcal{R}_0$ as given by the interleaver and code rate, making an illustration of this difficult. In addition, care must be taken not to include a given coupling in both $\mathcal{R}_1$ and $\mathcal{R}_{-1}$, resulting in a direct feedback-effect in the graph and thereby guaranteeing that the GBP algorithm will be non-convergent. Furthermore, as in the modified cluster variation method, marginals should be made consistent by requiring communication between regions containing the same variable. If required, this can be achieved by adding appropriate regions in e.g. $\mathcal{R}_{-1}$ and the region graph will therefore be given by the set of regions $\mathcal{R} = \mathcal{R}_0 \cup \mathcal{R}_1 \cup \mathcal{R}_{-1}$ with counting numbers given by (2.27). Using the notation

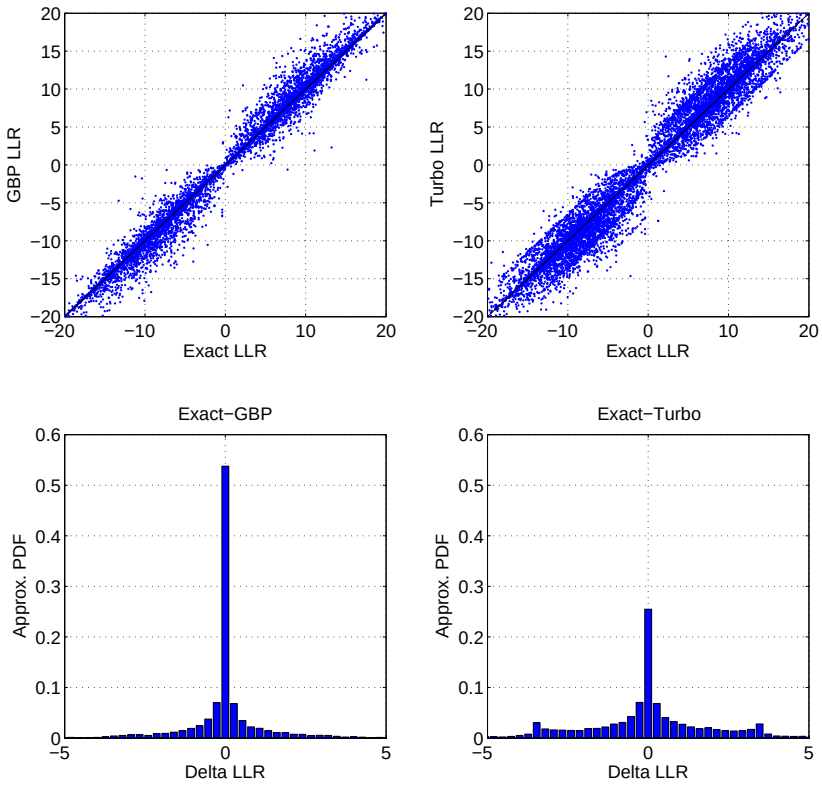


Figure 3.10: Comparison of traditional and generalized Turbo equalization, $SNR = 3dB$, $w_1 = 0.5$

in [YFW05], the resulting region graph is a junction graph with large regions $\mathcal{R}_L = \mathcal{R}_0$ and small regions $\mathcal{R}_S = \mathcal{R}_1 \cup \mathcal{R}_{-1}$. The counting numbers of regions in $\mathcal{R}_L$ must be one due to the fact that every region contains at least one factor node. The counting numbers of regions in $\mathcal{R}_S$ are given by $c_R = 1 - d_R$ where $d_R$ is the number of neighboring large regions, which for this type of junction graph is always 2, i.e. $c_R = -1$ for $R \in \mathcal{R}_S$.

As for the modified cluster variation method, significant book-keeping is required to construct the appropriate region graph for a given system, but as real-life systems are likely to use fixed coding schemes, the graph structure can be predetermined and put into a look-table. In figure 3.10, the LLRs achieved

for the same setup as in the modified cluster variation method is shown including an approximate distribution of the LLR errors from the GBP algorithm and the traditional Turbo principle. From this it can be seen that the generalized Turbo method provides similar results as achieved with the modified cluster variation method, but due to the longer loops in the region graph much less relaxation using the convex IIR method is required for convergence. In fact, convergence becomes less and less of a problem as the system size grows as is also the case for the traditional Turbo principle. The generalized Turbo principle was conceived very late in the research study and as a result of this, only the simulation in figure 3.10 is included here. However, the author firmly believes in the generality of this principle and hopes that this framework will provide a platform for future research.

## 3.6 Summary

This chapter has outlined several methods for approximate detection and decoding. First, the general form of linear detection is outlined as the full posterior provided by this may be of interest, e.g. for parameter estimation. Next, a practical method for performing whitening of noise/interference is described, providing flexibility and robustness to the remaining discrete signal set. Furthermore, the sphere detection and decoding framework is presented and a previously unknown connection between the QL factorization of the channel matrix and minimum-phase prefiltering is introduced. This establishes an unrecognized coupling between sphere detection and traditional RSSE, where sphere detection can be seen as a dynamic variant of RSSE with decisions taken only when a specified level of certainty has been achieved. Finally, the concept of using GBP on region graphs for approximate joint detection and decoding in systems with convolutional codes has been introduced. A modified cluster variation method is presented for this, but an excessive number of short loops in the graph makes convergence troublesome. As a result of this, the Turbo principle's method of generating long loops in the underlying graph is reused in a region graph setting. This gives rise to a generalized Turbo principle were region beliefs are exchanged between components instead of single-variable beliefs, as is the case for the traditional Turbo principle. Due to time constraints it has not been possible to fully investigate the proposed method by simulations before the thesis deadline, but a simple simulation is provided to indicate the improved performance offered by such an approach.

Chapter 4

# Parameter Estimation

For realistic communication systems, parameters assumed known until now must be estimated. In noisy channels, parameters can never be perfectly estimated and this chapter describes how parameter estimation in the generic system model can be integrated with the probabilistic detection and decoding framework presented in previous chapters. The content of this chapter is based primarily on the work published in [CL06, Chr07].

## 4.1 The Variational Bayesian EM Framework

The Variational Bayesian EM-algorithm (VBEM) is a recent [Mac97, Att00, BG03, Bea03] generalization of the classical EM-algorithm [DLR77, NH99] to a Bayesian setting. What this means is that we wish to estimate the posterior distribution of the model parameters instead of simply determining its maximum, as targeted by the EM-algorithm, and use this additional information to update the symbol posterior in the next iteration. Such an approach can have several beneficial properties as the estimated uncertainty of the parameter estimate is taken into account. Parameter estimation based on point-estimates, as performed by the EM-algorithm for e.g. channel estimation, seems fairly well-known in the communications society whereas generalizations of this taking uncertainty into account appear little known. In [TM03], the correct strategy

of averaging the likelihood function over the distribution of the channel estimate is considered, but the complexity of the resulting method seems unfeasible in general as the Markov structure of the channel is destroyed. The EM- and VBEM-algorithms avoid this problem by a bounding technique, which in turn averages over the log-likelihood function instead.

In a truly Bayesian framework, hidden variables and parameters are considered random variables with a given distribution that should be integrated out to produce the marginal likelihood of the model as

$$p\left(\mathbf{y}\right) = \int p\left(\mathbf{y}, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right) d\hat{\mathbf{x}} d\hat{\boldsymbol{\theta}} \tag{4.1}$$

with $\hat{\mathbf{x}}$ and $\hat{\boldsymbol{\theta}}$ being the unknown symbol and parameter set respectively. However, for most interesting models the integral in (4.1) is unfeasible and approximations must be made instead. Using Jensen's inequality, we can lower-bound the marginal log-likelihood as

$$\begin{aligned} ln\left[p\left(\mathbf{y}\right)\right] &= ln\left[\int q\left(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right) \frac{p\left(\mathbf{y}, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right)}{q\left(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right)} d\hat{\mathbf{x}} d\hat{\boldsymbol{\theta}}\right] \\ &\geq \int q\left(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right) ln\left[\frac{p\left(\mathbf{y}, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right)}{q\left(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right)}\right] d\hat{\mathbf{x}} d\hat{\boldsymbol{\theta}} \end{aligned} \tag{4.2}$$

where $q\left(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right)$ is a free distribution used to approximate the posterior $p\left(\mathbf{y}, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right)$. Maximizing the lower-bound w.r.t. the free distribution $q\left(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right)$ yields the exact posterior and determining this requires an exhaustive search over the joint distribution. A solution to this difficult problem is offered by constraining the free distribution to factorize in a convenient way and then maximize the lower-bound w.r.t. one component at a time. We know from section 3 how to compute the symbol posterior given the parameters, i.e. $p\left(\hat{\mathbf{x}} \mid \mathbf{y}, \hat{\boldsymbol{\theta}}\right)$, and classical estimation theory tells us[1] how to determine the parameter posterior given the transmitted symbols, i.e. $p\left(\hat{\boldsymbol{\theta}} \mid \mathbf{y}, \hat{\mathbf{x}}\right)$. A natural choice of factorization is therefore between the symbols and the parameters by requiring

$$q\left(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right) = q_{\hat{\mathbf{x}}}\left(\hat{\mathbf{x}}\right) q_{\hat{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}\right) \tag{4.3}$$

This factorization provides the intriguing solution of maximizing the lower-bound w.r.t. one of the distributions while fixing the other and vice versa. The

[1] The is only strictly true for noise known to be AWGN

individual maximization steps are given by the associated fixed-point equations found by taking the functional derivatives and equating to zero [Att00, Bea03]. The general result of this is that all but the component currently being updated should be integrated out in the log-domain as

$$\begin{aligned} VBE: \quad & q_{\hat{\mathbf{x}}}\left(\hat{\mathbf{x}}\right) \propto e^{\left\langle ln\left[p\left(\mathbf{y},\hat{\mathbf{x}}|\hat{\boldsymbol{\theta}}\right)\right]\right\rangle_{q_{\hat{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}\right)}} \\ VBM: \quad & q_{\hat{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}\right) \propto p\left(\hat{\boldsymbol{\theta}}\right) e^{\left\langle ln\left[p\left(\mathbf{y},\hat{\mathbf{x}}|\hat{\boldsymbol{\theta}}\right)\right]\right\rangle_{q_{\hat{\mathbf{x}}}\left(\hat{\mathbf{x}}\right)}} \end{aligned} \tag{4.4}$$

This effectively eliminates the notion of the E- and M-steps in the EM-algorithm as both are now expectations, but to maintain legacy the hidden variable and parameter updates are often called the VBE- and VBM-step respectively. If the parameter posterior $q_{\hat{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}\right)$ is further constrained to be a delta function, the EM-algorithm is recovered and the VBEM-algorithm can therefore be considered a generalization thereof.

From a communications point of view, this method boils down to performing the detection and decoding separately from the parameter estimation and then iterate between the two. A nice property of this approximation is that any detection and decoding scheme designed for known parameters can be utilized, but a potential problem is that the VBEM-algorithm can only guarantee convergence to a local maximum of $p\left(\mathbf{y}\right)$. Hence, there is a risk of getting stuck in a local maximum that may be far from optimal and initialization of the VBEM-algorithm is therefore equally important as for the EM-algorithm. In the communication systems of interest, this issue is handled by a training set, which hopefully starts the VBEM-algorithm sufficiently close to the optimum.

Until now, the VBEM-algorithm has been described in a very broad sense. Narrowing the focus to the system model of section 2.1, we have[2]

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon} \\ &= \mathbf{X}\mathbf{h} + \boldsymbol{\epsilon} \end{aligned} \tag{4.5}$$

with the lower equation being beneficial for channel estimation as the convolutional structure in $\mathbf{H}$ has been put into the symbol matrix $\mathbf{X}$ instead. In this manner, $\mathbf{h}$ fully describes $\mathbf{H}$ and similarly for $\mathbf{x}$ and $\mathbf{X}$. The true noise covariance is given by $\boldsymbol{\Sigma} \triangleq E\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^H\right]$ and the parameter set for this model is therefore $\boldsymbol{\theta} = \{\mathbf{h}, \boldsymbol{\Sigma}\}$. However, there is currently no known general non-iterative method of jointly estimating $\hat{\mathbf{h}}$ and $\hat{\boldsymbol{\Sigma}}$ for known symbols and the free distribution is therefore further constrained to factorize as

$$q\left(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\right) = q_{\hat{\mathbf{x}}}\left(\hat{\mathbf{x}}\right) q_{\hat{\mathbf{h}}}\left(\hat{\mathbf{h}}\right) q_{\hat{\boldsymbol{\Sigma}}}\left(\hat{\boldsymbol{\Sigma}}\right) \tag{4.6}$$

[2]The model used here is based on $\mathbf{y}$ as the whitened signal $\tilde{\mathbf{y}}$ is undefined when the noise covariance is unknown

The overall VBEM-algorithm for this system model will therefore consist of a

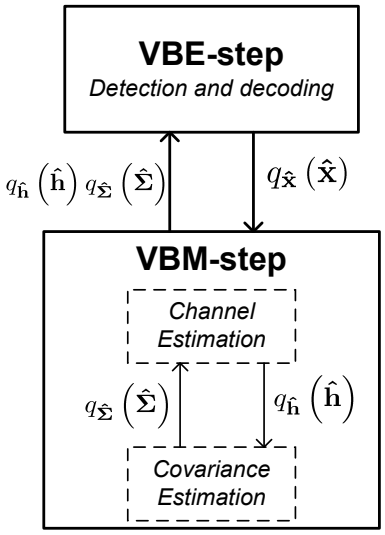


Figure 4.1: The individual components of the VBEM-algorithm for the system model.

VBM-step for parameter estimation, which consist of the channel and covariance estimators, and a VBE-step for symbol estimation. This separation into three separate components is illustrated in figure 4.1, where the distribution being updated is shown exiting the block updating it. The complexity of the VBEM-algorithm is comparable to that of the EM-algorithm, but tends to require more iterations to fully converge.

### 4.1.1 The VBM-step

First considering the channel estimate, it is shown in [CL06] that having the Gaussian prior $\mathbf{h} \sim \mathcal{CN}\left(\boldsymbol{\mu}_{\mathbf{h}}, \mathbf{\Sigma}_{\mathbf{h}}\right)$ results in a Gaussian posterior $\hat{\mathbf{h}} \sim q_{\hat{\mathbf{h}}}\left(\hat{\mathbf{h}}\right) = \mathcal{CN}\left(\boldsymbol{\mu}_{\hat{\mathbf{h}}}, \mathbf{\Sigma}_{\hat{\mathbf{h}}}\right)$ due to the prior being conjugate. The covariance and mean of this Gaussian posterior is given by

$$\begin{aligned} \mathbf{\Sigma}_{\hat{\mathbf{h}}} &= \left( \left\langle \hat{\mathbf{X}}^H \left\langle \hat{\mathbf{\Sigma}}^{-1} \right\rangle_{q_{\hat{\mathbf{\Sigma}}}\left(\hat{\mathbf{\Sigma}}\right)} \hat{\mathbf{X}} \right\rangle_{q_{\hat{\mathbf{x}}}\left(\hat{\mathbf{x}}\right)} + \mathbf{\Sigma}_{\mathbf{h}}^{-1} \right)^{-1} \\ \boldsymbol{\mu}_{\hat{\mathbf{h}}} &= \mathbf{\Sigma}_{\hat{\mathbf{h}}} \left( \left\langle \hat{\mathbf{X}} \right\rangle_{q_{\hat{\mathbf{x}}}\left(\hat{\mathbf{x}}\right)}^H \left\langle \hat{\mathbf{\Sigma}}^{-1} \right\rangle_{q_{\hat{\mathbf{\Sigma}}}\left(\hat{\mathbf{\Sigma}}\right)} \mathbf{y} + \mathbf{\Sigma}_{\mathbf{h}}^{-1} \boldsymbol{\mu}_{\mathbf{h}} \right) \end{aligned} \tag{4.7}$$

where, like for the linear detector in section 3.1, the mean $\boldsymbol{\mu}_{\hat{\mathbf{h}}}$ is both the MAP and MMSE channel estimate due to the factorization of (4.6). The posterior average $\left\langle \hat{\mathbf{X}} \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})}$ is simply the posterior mean of the symbols and it can therefore easily be found from $q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})$. The second-order moment in (4.7) is a little more complicated and will be discussed in section 4.1.3.

For ML noise covariance estimation, a necessary and sufficient statistic is the sample covariance matrix

$$\hat{\mathbf{S}}_{\hat{\boldsymbol{\epsilon}}} \triangleq \left\langle \hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^H \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})q_{\hat{\mathbf{h}}}(\hat{\mathbf{h}})} = \left\langle \left(\mathbf{y} - \hat{\mathbf{X}}\hat{\mathbf{h}}\right)\left(\mathbf{y} - \hat{\mathbf{X}}\hat{\mathbf{h}}\right)^H \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})q_{\hat{\mathbf{h}}}(\hat{\mathbf{h}})} =$$
$$\mathbf{y}\mathbf{y}^H - \mathbf{y}\boldsymbol{\mu}_{\hat{\mathbf{h}}}^H \left\langle \hat{\mathbf{X}} \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})}^H - \left(\mathbf{y}\boldsymbol{\mu}_{\hat{\mathbf{h}}}^H \left\langle \hat{\mathbf{X}} \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})}^H\right)^H + \left\langle \hat{\mathbf{X}}^H \left(\boldsymbol{\mu}_{\hat{\mathbf{h}}}\boldsymbol{\mu}_{\hat{\mathbf{h}}}^H + \boldsymbol{\Sigma}_{\hat{\mathbf{h}}}\right)\hat{\mathbf{X}} \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})} \tag{4.8}$$

where the following relations for the Gaussian channel estimate have been exploited in (4.8)

$$\begin{aligned} \left\langle \hat{\mathbf{h}} \right\rangle_{q_{\hat{\mathbf{h}}}(\hat{\mathbf{h}})} &= \boldsymbol{\mu}_{\hat{\mathbf{h}}} \\ \left\langle \hat{\mathbf{h}}\hat{\mathbf{h}}^H \right\rangle_{q_{\hat{\mathbf{h}}}(\hat{\mathbf{h}})} &= \boldsymbol{\mu}_{\hat{\mathbf{h}}}\boldsymbol{\mu}_{\hat{\mathbf{h}}}^H + \boldsymbol{\Sigma}_{\hat{\mathbf{h}}} \end{aligned} \tag{4.9}$$

As mentioned in section 3.2 in connection with whitening-based detection, a major problem with the formulation of (4.8) is that it may be ill-conditioned, possibly even rank-one. In [Chr05b], this problem is circumvented by breaking the covariance down as described in section 3.2, but another and potentially better method is to exploit the known structure of the noise covariance to provide a constrained estimate. Such a method capable of operating directly on (4.8) will be described in section 4.2.

Due to the inverse-Wishart conjugate prior $\boldsymbol{\Sigma} \sim \mathcal{CW}^{-1}(N_\nu, N_\nu \mathbf{S}_{\boldsymbol{\Sigma}})$ applied for the noise covariance matrix, the required moment of the noise covariance posterior is given by[3]

$$\left\langle \hat{\boldsymbol{\Sigma}}^{-1} \right\rangle_{q_{\hat{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{\Sigma}})} = (N_o + N_\nu)\left(N_o \hat{\mathbf{S}}_{\hat{\boldsymbol{\epsilon}}} + N_\nu \mathbf{S}_{\boldsymbol{\Sigma}}\right)^{-1} \tag{4.10}$$

and the inverse-Wishart conjugate prior can therefore be interpreted as inserting $N_\nu$ virtual noise observations $\{\boldsymbol{\epsilon}_\nu\}_{\nu=1}^{N_\nu}$ having sample covariance $\mathbf{S}_{\boldsymbol{\Sigma}} \triangleq N_\nu^{-1}\sum_{\nu=1}^{N_\nu}\boldsymbol{\epsilon}_\nu\boldsymbol{\epsilon}_\nu^H$ with $N_\nu = 0$ recovering the ML solution. As a result of this, it is only the channel estimate that needs to have a full distribution as only (4.10)

[3] Here the generalization to $N_o$ independent vector observations is considered

is required to produce the whitened signal $\tilde{\mathbf{y}} \triangleq \hat{\mathbf{F}}\mathbf{y}$ where $\hat{\mathbf{F}}$ is the Cholesky factor of (4.10). The VBE-step is therefore expressed in terms of the whitened system model $\tilde{\mathbf{y}} \triangleq \hat{\mathbf{F}}\hat{\mathbf{X}}\hat{\mathbf{h}} + \hat{\mathbf{F}}\hat{\boldsymbol{\epsilon}} = \hat{\mathbf{F}}\hat{\mathbf{H}}\hat{\mathbf{x}} + \hat{\mathbf{F}}\hat{\boldsymbol{\epsilon}}$ as this eliminates the need for a full noise covariance distribution.

### 4.1.2 The VBE-step

The only missing component of the VBEM-algorithm is now the symbol estimation, which is traditionally performed by detection and decoding schemes such as those described in chapter 3, but the VBM-step provides a Gaussian channel distribution and this should be exploited. For all detectors, the symbol posterior coming from the VBE-step can be expressed as

$$q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}) \propto e^{-\left\langle f\left(\tilde{\mathbf{y}},\hat{\mathbf{x}},\hat{\mathbf{F}},\hat{\mathbf{h}}\right)\right\rangle_{q_{\hat{\mathbf{h}}}(\hat{\mathbf{h}})} + ln[p(\hat{\mathbf{x}})]} \tag{4.11}$$

where $f\left(\tilde{\mathbf{y}},\hat{\mathbf{x}},\hat{\mathbf{F}},\hat{\mathbf{h}}\right)$ is a function specific to the detector used and $p(\hat{\mathbf{x}})$ is the symbol prior. Most detectors rely on a negative log-likelihood term similar in structure to that used in optimal linear and non-linear detection, i.e.

$$f\left(\tilde{\mathbf{y}},\hat{\mathbf{x}},\hat{\mathbf{F}},\hat{\mathbf{h}}\right) = \|\tilde{\mathbf{y}} - \hat{\mathbf{F}}\hat{\mathbf{X}}\hat{\mathbf{h}}\|^2 \tag{4.12}$$

but some detectors may e.g. brake it up into smaller independent sums resulting in approximate detection.

The interesting point here is, that the log-likelihood function should be averaged over the distribution of the channel estimate yielding

$$\left\langle f\left(\tilde{\mathbf{y}},\hat{\mathbf{x}},\hat{\mathbf{F}},\hat{\mathbf{h}}\right)\right\rangle_{q_{\hat{\mathbf{h}}}(\hat{\mathbf{h}})} = \underbrace{\|\tilde{\mathbf{y}} - \hat{\mathbf{F}}\hat{\mathbf{X}}\boldsymbol{\mu}_{\hat{\mathbf{h}}}\|^2}_{EM} + \underbrace{tr\left\{\hat{\mathbf{X}}^H\hat{\mathbf{F}}^H\hat{\mathbf{F}}\hat{\mathbf{X}}\boldsymbol{\Sigma}_{\hat{\mathbf{h}}}\right\}}_{Penalty} \tag{4.13}$$

with the penalty term being a positive quantity added due to the uncertainty of the channel estimate. The expression in (4.13) can be used directly in detectors based on discrete state-space models, such as the FBA, as modified metrics.

Considering detectors based on Gaussian symbol assumptions, the matrix channel formulation of the negative log-likelihood is desired, resulting in

$$\left\langle f\left(\tilde{\mathbf{y}},\hat{\mathbf{x}},\hat{\mathbf{F}},\hat{\mathbf{h}}\right)\right\rangle_{q_{\hat{\mathbf{h}}}(\hat{\mathbf{h}})} = \underbrace{\|\tilde{\mathbf{y}} - \hat{\mathbf{F}}\boldsymbol{\mu}_{\hat{\mathbf{H}}}\hat{\mathbf{x}}\|^2}_{EM} + \underbrace{\hat{\mathbf{x}}^H\boldsymbol{\Sigma}_{\hat{\mathbf{H}},\hat{\mathbf{F}}^H\hat{\mathbf{F}}}\hat{\mathbf{x}}}_{Penalty} \tag{4.14}$$

Here, $\boldsymbol{\mu}_{\hat{\mathbf{H}}} \triangleq \left\langle \hat{\mathbf{H}} \right\rangle_{q_{\hat{\mathbf{h}}}(\hat{\mathbf{h}})}$ is the posterior mean of the channel matrix and $\boldsymbol{\Sigma}_{\hat{\mathbf{H}},\hat{\mathbf{F}}^H\hat{\mathbf{F}}}$ is the covariance of the channel matrix defined in section 4.1.3. Due to the modification of the log-likelihood function, the resulting Gaussian symbol posterior

$q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}) = \mathcal{CN}(\boldsymbol{\mu}_{\hat{\mathbf{x}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{x}}})$, traditionally given by (3.3), is now specified by

$$\begin{aligned} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}} &= \left(\boldsymbol{\mu}_{\hat{\mathbf{H}}}^H \hat{\mathbf{F}}^H \hat{\mathbf{F}} \boldsymbol{\mu}_{\hat{\mathbf{H}}} + \boldsymbol{\Sigma}_{\hat{\mathbf{H}},\hat{\mathbf{F}}^H\hat{\mathbf{F}}} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\right)^{-1} \\ \boldsymbol{\mu}_{\hat{\boldsymbol{x}}} &= \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}} \left(\boldsymbol{\mu}_{\hat{\mathbf{H}}}^H \hat{\mathbf{F}}^H \hat{\mathbf{F}} \mathbf{y} + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}}\right) \end{aligned} \tag{4.15}$$

As a result of this, the modified MMSE detector given by (4.15) accounts for the uncertainty in the channel estimate in much the same way as a zero-mean prior is taken into account.

### 4.1.3 Posterior Second-Order Moments

In the VBEM-algorithm, the use of posterior second-order moments are common, i.e. terms such as

$$\begin{aligned} &\left\langle \hat{\mathbf{X}}^H \mathbf{A} \hat{\mathbf{X}} \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})} \\ &\left\langle \hat{\mathbf{H}}^H \mathbf{A} \hat{\mathbf{H}} \right\rangle_{q_{\hat{\mathbf{h}}}(\hat{\mathbf{h}})} \end{aligned} \tag{4.16}$$

Here $\mathbf{A}$ is an arbitrary covariance-like matrix and matrices $\hat{\mathbf{X}}$ and $\hat{\mathbf{H}}$ are structured convolution matrices with the upper term of (4.16) being present in the VBM-step and the lower in the VBE-step, if linear detection is used. As the two share the same structure, only the top one will be examined, but results will also apply to the lower one.

Defining the posterior mean of the matrix as $\boldsymbol{\mu}_{\hat{\mathbf{X}}} \triangleq \left\langle \hat{\mathbf{X}} \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})}$ makes it possible to rewrite (4.16) as

$$\left\langle \hat{\mathbf{X}}^H \mathbf{A} \hat{\mathbf{X}} \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})} = \boldsymbol{\mu}_{\hat{\mathbf{X}}}^H \mathbf{A} \boldsymbol{\mu}_{\hat{\mathbf{X}}} + \underbrace{\left\langle \left(\hat{\mathbf{X}} - \boldsymbol{\mu}_{\hat{\mathbf{X}}}\right)^H \mathbf{A} \left(\hat{\mathbf{X}} - \boldsymbol{\mu}_{\hat{\mathbf{X}}}\right) \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})}}_{\boldsymbol{\Sigma}_{\hat{\mathbf{X}},\mathbf{A}}} \tag{4.17}$$

The left-most part of (4.17) comes from the point-estimate $\boldsymbol{\mu}_{\hat{\mathbf{X}}}$ and the right-most term $\boldsymbol{\Sigma}_{\hat{\mathbf{X}},\mathbf{A}}$ is the covariance expressing the uncertainty of the estimate in the space spanned by $\mathbf{A}$. The covariance matrix $\boldsymbol{\Sigma}_{\hat{\mathbf{X}},\mathbf{A}}$ can be constructed from linear combinations of scalar covariances having the form

$$\sigma_{i,j} \triangleq \left\langle \left(\hat{x}_i - \mu_{\hat{x}_i}\right) \left(\hat{x}_j - \mu_{\hat{x}_j}\right)^* \right\rangle_{q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})} \tag{4.18}$$

If $q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}) = \mathcal{CN}(\boldsymbol{\mu}_{\hat{\mathbf{x}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{x}}})$, then clearly $\sigma_{i,j} = [\boldsymbol{\Sigma}_{\hat{\mathbf{x}}}]_{i,j}$. On the other hand, if $q_{\hat{\mathbf{x}}}(\hat{\mathbf{x}})$ is discrete with $x_i \in \Omega$, then computing the covariance in (4.18) consists of averaging over the $|\Omega|^2$ possible values of the posterior distribution.

A point that should be stressed here is the option of using marginal posteriors in (4.18), i.e. fully factorized posteriors of the form

$$q_{\hat{\mathbf{x}}}\left(\hat{\mathbf{x}}\right) = \prod_i q_{\hat{x}_i}\left(\hat{x}_i\right) \tag{4.19}$$

Using this approximation simplifies (4.18) to

$$\sigma_{i,j} = \left\langle \left|\hat{x}_i - \mu_{\hat{x}_i}\right|^2 \right\rangle_{q_{\hat{x}_i}(\hat{x}_i)} \qquad , i = j \tag{4.20}$$

and zero for $i \neq j$. This method of using marginals seems to be prevalent in the communications society as it simplifies the design (see e.g. [OT04] and references therein), but it does not account for the full uncertainty in the estimate and better results can therefore be expected by exploiting full posteriors instead of only marginals. Detectors derived based on the full factorization in (4.19) can naturally not capture the full posterior and various correction schemes exist trying to improve on this, see e.g. [FN02] and references therein.

### 4.1.4 Comparison of VBEM and EM for Parameter Estimation

To indicate the advantage of the VBEM-algorithm over the EM-algorithm, a single-user uncoded GSM-like system was simulated in [CL06] using BPSK modulation and the FBA for detection. In figure 4.2, the BER of a system using the GSM frame format of 142 information bits and 3 tail bits at each end of the frame is shown. However, according to the GSM frame format, the center of every frame contains a training set of 26 known bits, which is used to initialize the iterative algorithms. The task of the EM- or VBEM-algorithm is now to estimate both the channel and the noise power. For reference, the BER using known parameters and parameters estimated using ML estimation from the training set only is also shown.

In the right of the figure, it can be seen that there is virtually no difference between EM and VBEM. The reason is that the frame is large enough to make the posteriors highly peaked around the ML solution. On the other hand, if the frame length and training set is reduced to half of that specified by the GSM specification, a gain of about 0.3dB is achieved by VBEM compared to EM as the posteriors are now less peaked. Hence, the VBEM-algorithm is only advantageous when the individual posteriors of the system have significant variations around their ML/MAP solutions. If this is not the case, the VBEM-algorithm simply falls back to the associated EM-algorithm. However, for coded communication systems with many parameters to estimate from a limited set of observations, the VBEM-algorithm should prove beneficial.

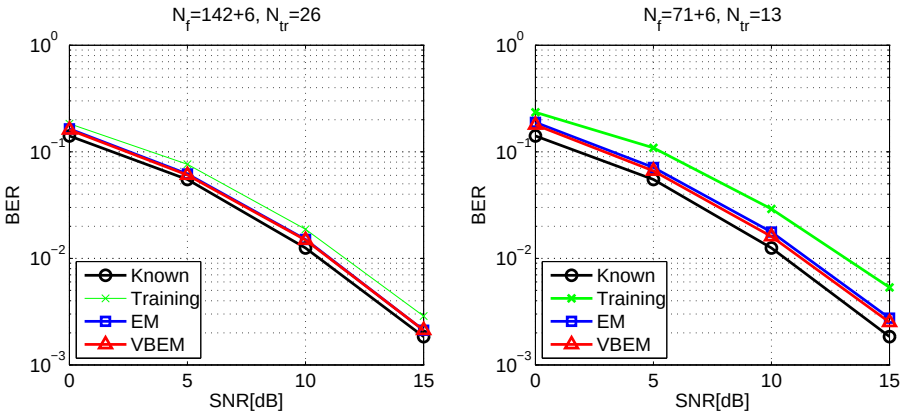


Figure 4.2: GSM-like system in AWGN using a TUx channel model and non-informative priors for parameter estimation, L=7, K'=K=1.

## 4.2 Band-Toeplitz Covariance Estimation

As mentioned previously in connection with whitening-based detection and the VBEM-algorithm, estimating the full covariance matrix is difficult. This is a result of the raw sample covariance matrix, which is also the unstructured ML estimate, being a low-rank estimate, possibly even rank one. However, due to the overall noise consisting of interference and filtered AWGN as given by (2.3), the noise covariance matrix can be seen to have a block-banded block-Toeplitz structure. In general, if a Toeplitz matrix has non-zero elements only up to and including lag $B$, it is called band-Toeplitz with bandwidth $B$. As the overall noise consists of interferers with a channel of length $L$, the bandwidth of the noise covariance matrix is $L-1$. Requiring that any estimate should have this structure results in fewer degrees-of-freedom and thus a better determined estimate. The problem of this idea is that the estimation becomes more difficult as the elements of the structured covariance estimate are now coupled. In the following, only a block-size of one is considered, i.e. band-Toeplitz covariance estimation. However, all methods presented naturally generalize to larger block-sizes.

A natural solution to this problem of ML band-Toeplitz covariance estimation is to compute the gradient of the log-likelihood function and requiring it to be zero for the structured estimate. Let $\mathbf{S}_{\boldsymbol{\epsilon}} \triangleq N_o^{-1} \sum_{i=1}^{N_o} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^H$ be the sample covariance of $N_o$ independent zero-mean complex-valued Gaussian vector observations $\mathcal{E} \triangleq \{\boldsymbol{\epsilon}_i\}_{i=1}^{N_o}$ with $\boldsymbol{\epsilon}_i \in \mathbb{C}^M$. The log-likelihood of the covariance estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is then

$$\mathcal{L}\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}\right) \triangleq ln\left[p\left(\mathcal{E} \mid \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}\right)\right] = -N_o ln\left[\left|\pi\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}\right|\right] - N_o tr\left\{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\mathbf{S}_{\boldsymbol{\epsilon}}\right\} \tag{4.21}$$

and maximizing (4.21) under the structural constraints on $\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}$ will provide a (local) ML estimate with the desired properties. To do this, the first-order derivative is easily found to be

$$\frac{\partial \mathcal{L}\left(\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}\right)}{\partial \hat{c}_b} = N_o tr\left\{\frac{\partial \hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}}{\partial \hat{c}_b}\left(\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\mathbf{S}_{\boldsymbol{\epsilon}}\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} - \hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\right)\right\} \tag{4.22}$$

where $\hat{c}_b$ is the lag $b$ correlation coefficient in $\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}$. It is now desired to find a set of coefficients $\hat{\mathbf{c}} \triangleq [\hat{c}_0, \cdots, \hat{c}_B]^T$ such that (4.22) is zero for all of these $B+1$ coefficients. Unfortunately, the inverses in (4.22) constitute a highly non-linear system of equations in $\hat{\mathbf{c}}$ and solving it requires advanced optimization schemes. However, the perhaps biggest obstacle is the constraint of non-negative definiteness required for all valid covariance matrices and this additional constraint must be incorporated into the optimization. In general, it is therefore believed that such methods will be inferior to band-Toeplitz covariance estimation based on the EM-algorithm, as will be described next.

### 4.2.1 EM-based Toeplitz Covariance Estimation

In [MS87, DMS89, FB90], an EM-algorithm is formulated based on a circulant extension of the Toeplitz matrix. Applying this idea to the VBEM framework, the equivalent of (4.1) for this system becomes[4]

$$p\left(\boldsymbol{\epsilon}\right) = \int p\left(\boldsymbol{\epsilon}, \hat{\mathbf{f}}, \hat{\boldsymbol{\theta}}\right) d\hat{\mathbf{f}} d\hat{\boldsymbol{\theta}} \tag{4.23}$$

Here $\hat{\mathbf{f}} \in \mathbb{C}^{M_z - M}$ are hypothesized observations that make estimation easier, similar to how transmitted symbols are hypothesized in (4.1). For a circulant extension to be possible, $M_z \geq 2M-1$ and the complete-data $\hat{\mathbf{z}} \triangleq \left[\boldsymbol{\epsilon}^T, \hat{\mathbf{f}}^T\right]^T$ must be restricted to being jointly Gaussian with parameter set $\hat{\boldsymbol{\theta}}$. There is therefore a many-to-one mapping from the complete-data $\hat{\mathbf{z}}$ to the incomplete-data $\boldsymbol{\epsilon}$ given by

$$\boldsymbol{\epsilon} = \underbrace{\left[\mathbf{I}_M, \mathbf{0}_{M\times(M_z-M)}\right]}_{\mathbf{P}}\hat{\mathbf{z}} \tag{4.24}$$

As samples are assumed to be zero-mean, only the covariance of $\hat{\mathbf{z}}$ is required to describe the distribution, i.e. $\hat{\boldsymbol{\theta}} = \left\{\hat{\mathbf{\Sigma}}_{\hat{\mathbf{z}}}\right\}$ where $\hat{\mathbf{\Sigma}}_{\hat{\mathbf{z}}} \in \mathbb{C}^{M_z \times M_z}$ is constrained to being circulant.

[4] For now, $N_o = 1$ is assumed to simplify expressions

Like in (4.2), we lower-bound the marginal log-likelihood by Jensen's inequality as

$$ln\left[p\left(\boldsymbol{\epsilon}\right)\right] \geq \int q\left(\hat{\mathbf{f}}, \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right) ln\left[\frac{p\left(\boldsymbol{\epsilon}, \hat{\mathbf{f}}, \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right)}{q\left(\hat{\mathbf{f}}, \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right)}\right] d\hat{\mathbf{f}} d\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}} \tag{4.25}$$

As before, the free distribution is then constrained to factorize between the hidden variables and the parameters as

$$\begin{aligned} q\left(\hat{\mathbf{f}}, \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right) &= q_{\hat{\mathbf{f}}}\left(\hat{\mathbf{f}}\right) q_{\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}}\left(\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right) \\ &= q_{\hat{\mathbf{z}}}\left(\hat{\mathbf{z}}\right) q_{\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}}\left(\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right) \end{aligned} \tag{4.26}$$

where the last line is due to $\hat{\mathbf{z}}$ uniquely determining $\boldsymbol{\epsilon}$. The distribution $q_{\hat{\mathbf{z}}}\left(\hat{\mathbf{z}}\right)$ is therefore a conditional distribution resulting in the updates being

$$\begin{aligned} VBE: \quad & q_{\hat{\mathbf{z}}}\left(\hat{\mathbf{z}}\right) \propto e^{\left\langle ln\left[p\left(\hat{\mathbf{z}}|\boldsymbol{\epsilon}, \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right)\right]\right\rangle_{q_{\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}}\left(\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right)}} \\ VBM: \quad & q_{\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}}\left(\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right) \propto e^{\left\langle ln\left[p\left(\hat{\mathbf{z}}|\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right)\right]\right\rangle_{q_{\hat{\mathbf{z}}}\left(\hat{\mathbf{z}}\right)}} \propto \left|\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right|^{-1} e^{-tr\left\{\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}^{-1}\left\langle \hat{\mathbf{z}}\hat{\mathbf{z}}^H\right\rangle_{q_{\hat{\mathbf{z}}}\left(\hat{\mathbf{z}}\right)}\right\}} \end{aligned} \tag{4.27}$$

As for general covariance estimation described in section 4.1.1, only the MAP estimate

$$\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m} \triangleq \arg\max_{\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}} q_{\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}}\left(\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right) \tag{4.28}$$

is required in the VBE-step. This is a result of the inverse-Wishart posterior having the property $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m}^{-1} \propto \left\langle \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}^{-1} \right\rangle_{q_{\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}}\left(\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}\right)}$ and the VBEM-algorithm therefore falls back to being an EM-algorithm as effectively only a point-estimate is used. A conjugate prior on the covariance estimate has been excluded here for simplicity, but may be incorporated as in (4.10) if desired.

Using properties of conditional Gaussian distributions, the E-step can be found to be

$$q_{\hat{\mathbf{z}}}\left(\hat{\mathbf{z}}\right) = \mathcal{CN}\left(\boldsymbol{\mu}_{\hat{\mathbf{z}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{z}}}\right) \tag{4.29}$$

with

$$\begin{aligned} \boldsymbol{\mu}_{\hat{\mathbf{z}}} &= \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m} \mathbf{P}^H \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m}^{-1} \boldsymbol{\epsilon} \\ \boldsymbol{\Sigma}_{\hat{\mathbf{z}}} &= \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m} - \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m} \mathbf{P}^H \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m}^{-1} \mathbf{P} \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m} \end{aligned} \tag{4.30}$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m}$ is the upper-left sub-matrix of $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m}$, i.e.

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m} \triangleq \mathbf{P} \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m} \mathbf{P}^H \tag{4.31}$$

The sample covariance of the complete-data required in the M-step is then given by

$$\begin{aligned}\hat{\mathbf{S}}_{\hat{\mathbf{z}}} &\triangleq \left\langle \hat{\mathbf{z}}\hat{\mathbf{z}}^H \right\rangle_{q_{\hat{\mathbf{z}}}(\hat{\mathbf{z}})} = \boldsymbol{\mu}_{\hat{\mathbf{z}}}\boldsymbol{\mu}_{\hat{\mathbf{z}}}^H + \boldsymbol{\Sigma}_{\hat{\mathbf{z}}} \\ &= \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m} + \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m}\mathbf{P}^H \left( \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m}^{-1}\mathbf{S}_{\boldsymbol{\epsilon}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m}^{-1} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m}^{-1} \right) \mathbf{P}\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}},m}\end{aligned} \tag{4.32}$$

where $\boldsymbol{\epsilon}\boldsymbol{\epsilon}^H$ has been replaced by $\mathbf{S}_{\boldsymbol{\epsilon}}$ to generalize the expression beyond $N_o = 1$. We now have an EM-algorithm consisting of an E-step given by (4.32) and a M-step given by (4.27) and (4.28).

However, the whole idea of using an EM-algorithm was to produce the complete-data sample covariance so that Toeplitz estimation would be easy by requiring the complete-data covariance estimate to be circulant. The maximization in (4.28) should therefore only be performed over the space of circulant matrices, but as it is well-known that the Discrete Fourier Transform (DFT) matrix diagonalizes any circulant matrix, this is easy. Letting $\mathbf{D} \in \mathbb{C}^{M_z \times M_z}$ be the normalized DFT matrix, then

$$\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}} = \mathbf{D}\hat{\boldsymbol{\Lambda}}_{\hat{\mathbf{z}}}\mathbf{D}^H \tag{4.33}$$

where the diagonal matrix $\hat{\boldsymbol{\Lambda}}_{\hat{\mathbf{z}}} \triangleq diag\left(\hat{\boldsymbol{\lambda}}_{\hat{\mathbf{z}}}\right)$ contains the eigenvalues of $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{z}}}$. Maximizing (4.28) over the space of circulant matrices is therefore done by maximizing over it's eigenvalues, which is easily found to be

$$\hat{\boldsymbol{\lambda}}_{\hat{\mathbf{z}}} = diag\left(\mathbf{D}^H\mathbf{S}_{\hat{\mathbf{z}}}\mathbf{D}\right) \tag{4.34}$$

where $diag\left(\cdot\right)$ extracts the diagonal from the matrix. To summarize the EM-algorithm including the iteration index $j$, the algorithm can be expressed as

$$\begin{aligned} E: \quad & \boldsymbol{\Delta}^{(j)} = \hat{\boldsymbol{\Lambda}}_{\hat{\mathbf{z}}}^{(j)}\tilde{\mathbf{D}}\left(\mathbf{W}^{(j)}\mathbf{S}_{\boldsymbol{\epsilon}}\mathbf{W}^{(j)} - \mathbf{W}^{(j)}\right)\tilde{\mathbf{D}}^H\hat{\boldsymbol{\Lambda}}_{\hat{\mathbf{z}}}^{(j)} \\ M: \quad & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m}^{(j+1)} = \tilde{\mathbf{D}}^H diag\left(\underbrace{\hat{\boldsymbol{\lambda}}^{(j)} + diag\left(\boldsymbol{\Delta}^{(j)}\right)}_{\hat{\boldsymbol{\lambda}}^{(j+1)}}\right)\tilde{\mathbf{D}}\end{aligned} \tag{4.35}$$

with $\tilde{\mathbf{D}} \triangleq \mathbf{D}\mathbf{P}^H$, $\mathbf{W}^{(j)} \triangleq \left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m}^{(j)}\right)^{-1}$ and $\boldsymbol{\Delta}^{(j)}$ being the unconstrained update to $\mathbf{S}_{\hat{\mathbf{z}}}$. A nice feature of this EM-algorithm is that the constraint of the estimate being non-negative definite is naturally fulfilled as can be seen from (4.34). Further, as the rank of $\mathbf{S}_{\boldsymbol{\epsilon}}$ is typically low, an efficient E-step can be implemented by FFTs and the Toeplitz structure can be exploited for computing $\mathbf{W}^{(j)}$ [DMS89]. As for any EM-algorithm, initialization is crucial and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},m}^{(0)} = N_o^{-1}tr\left\{\mathbf{S}_{\boldsymbol{\epsilon}}\right\}\mathbf{I}_M$ seems to be a reasonable choice.

### 4.2.2 Extension to Band-Toeplitz Covariance Estimation

In [Chr07], the EM-algorithm is extended to band-Toeplitz estimation by further constraining the M-step and leaving the E-step unchanged as given by (4.35). The basic idea is then, that if the bandwidth of the desired estimate is known to be $B$, the eigenvalues $\hat{\boldsymbol{\lambda}}_{\hat{\mathbf{z}}}$ may span only the lowest $B$ "frequencies" of $\mathbf{D}$, i.e.

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{\hat{\mathbf{z}}} &= M_z^{\frac{1}{2}} \sum_{b=-B}^{B} \hat{c}_b \mathbf{d}_b \qquad s.t. \quad \hat{c}_b = \hat{c}_{-b}^* \\ &= M_z^{\frac{1}{2}} \left( \hat{c}_0 \mathbf{d}_0 + 2 \sum_{b=1}^{B} Re\left\{ \hat{c}_b \mathbf{d}_b \right\} \right)\end{aligned} \tag{4.36}$$

where $\mathbf{d}_b = \mathbf{d}_{-b}^* \triangleq [\mathbf{D}]_{:,mod(b,M_z)}$. However, as the eigenvalues are real-valued and there are only $N_b \triangleq 2B+1$ real-valued degrees-of-freedom in the constrained estimate, we choose to reformulate the constraint as a real-valued decomposition, i.e.

$$\hat{\boldsymbol{\lambda}}_{\hat{\mathbf{z}}} = \mathbf{T}\hat{\mathbf{r}} \tag{4.37}$$

Here $\hat{\mathbf{r}} \in \mathbb{R}^{N_b}$ are the unknowns and $\mathbf{T} \triangleq [\mathbf{t}_0, \ldots, \mathbf{t}_{N_b-1}]$ is defined by

$$\mathbf{t}_b \triangleq \begin{cases} M_z^{\frac{1}{2}} Re\left\{ \mathbf{d}_b \right\} & , b = 0 \\ 2M_z^{\frac{1}{2}} Re\left\{ \mathbf{d}_b \right\} & , 1 \leq b \leq B \\ -2M_z^{\frac{1}{2}} Im\left\{ \mathbf{d}_{b-B} \right\} & , B < b < N_b \end{cases} \tag{4.38}$$

The upper $B+1$ coefficients of $\hat{\mathbf{r}}$ will therefore hold the real part of $\hat{\mathbf{c}}$ while the lower $B$ coefficients are the imaginary part. In the case of real-valued covariance estimation, having $N_b \triangleq B+1$ is therefore sufficient to parameterize the constrained estimate. The challenge is now, given the current ML estimate over the space of circulant matrices $\hat{\boldsymbol{\lambda}}_{\hat{\mathbf{z}}}$, to minimize some distance measure $f\left(\hat{\boldsymbol{\lambda}}_{\hat{\mathbf{z}}}, \hat{\mathbf{r}}\right)$ between the Toeplitz and the band-Toeplitz estimate subject to the constraint that the resulting eigenvalues must all be non-negative, i.e. $\mathbf{Tr} \geq \mathbf{0}_{M_z \times 1}$.

The distance measure $f\left(\hat{\boldsymbol{\lambda}}_{\hat{\mathbf{z}}}, \hat{\mathbf{r}}\right)$ required for ML estimation is easily found by inserting (4.37) into the M-step and is given in [Chr07] along with its derivatives required for minimization. Furthermore, in [Chr07] the Unbiased Cramer-Rao Lower-Bound (U-CRLB) on the variance of band-Toeplitz covariance estimation is derived for comparison. Also, the Weighted Projected (WP) method of Toeplitz covariance estimation is presented, which is a simple linear estimator that straight-forwardly integrates with the whitening-based detector presented in section 3.2. Basically, WP can be seen as nothing more than the covariance matrix generalization of classical correlation-based power-spectrum estimation

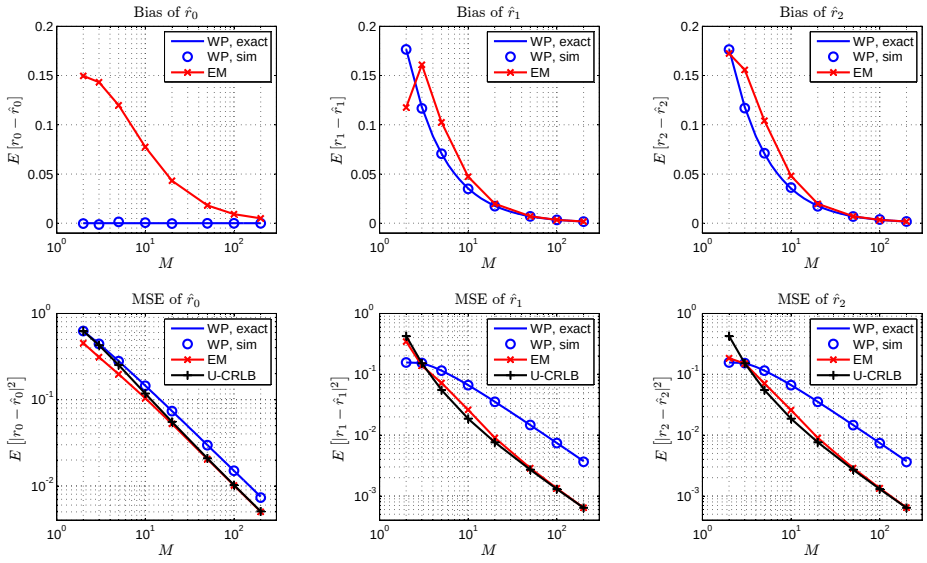


Figure 4.3: Example of band-Toeplitz estimation for varying number of observations $M$ for $B = 1$, $N_o = 1$.

employing triangular windowing to guarantee a positive power-spectrum, see e.g. [Hay96].

In figure 4.3, the WP estimator is compared with the proposed EM-based estimator for estimation of correlations in AWGN filtered by a first-order filter having coefficients $\left[\frac{1}{\sqrt{2}}, \frac{1+\sqrt{-1}}{2}\right]^T$, i.e. having bandwith $B = 1$, as a function of the number of observations $M$. The true correlations are therefore given by $\mathbf{c} = \left[1, \frac{1+\sqrt{-1}}{\sqrt{8}}\right]^T$ or equivalently $\mathbf{r} = \left[1, \frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}\right]^T$. From the figure, it should be noticed how a significant reduction in MSE can be achieved with only a small increase in bias and how the estimator approaches the U-CRLB for medium and large sample-sizes. The case of $M = 2$ is a special case as there is no bandwidth constraint and the proposed EM-algorithm therefore falls back to Toeplitz estimation. This is believed to be the reason for the apparent discontinuity at that point for the proposed EM-algorithm. As a result of the added bandwidth constraint, the number of iterations required for convergence is significantly reduced compared to that of Toeplitz estimation. For the above example, approximately 5 iterations seem sufficient whereas upwards of 50 iterations is required for Toeplitz estimation in the same system. Overall, the proposed estimator is believed to achieve near-optimal performance for band-Toeplitz covariance estimation while preserving a reasonable complexity through the use of FFTs and efficient computations on band-Toeplitz systems of equations [DMS89]. An

interesting option that has not been investigated further is the option of modifying the proposed method for constraining not the covariance itself, but its inverse to being band-Toeplitz. This can be achieved by constraining not the eigenvalues of the circulant covariance itself, but instead applying the same constraint on its inverse eigenvalues. Effectively, this is equivalent to the estimation of a stationary Gauss-Markov process of a specified Markov order and such an estimate would also fit nicely with the whitening framework.

## 4.3 Summary

First, this chapter has outlined the VBEM-algorithm, which is a general framework for separate detection/decoding and parameter estimation. The framework generalizes that of the EM-algorithm by exploiting the parameter posterior distribution for detection/decoding instead of the MAP estimate. However, the VBEM-algorithm is only beneficial if the parameter posterior is less peaked around the MAP estimate as it otherwise simply falls back to being an EM-algorithm. A point not discussed by this thesis is model selection, i.e. selection of parameters such as channel length etc. However, this has a natural integration with the VBEM framework by considering the model yet another unknown that should be integrated out. Doing so has the benefit that the number of parameters in the considered models are taken into account, making such methods less susceptible to over-fitting, see e.g. [Bea03]. From a real-life implementation point-of-view, this makes the VBEM-algorithm even more interesting, but it also has a higher complexity as parameter estimation must be performed in a number of models instead of just one.

Furthermore, an EM-algorithm for band-Toeplitz covariance estimation has been described as a generalization of an existing EM-algorithm for Toeplitz covariance estimation. The proposed method works by modifying the existing M-step so as to guarantee the desired bandwidth constraint of the estimate. Comparing performance of the proposed method with the unbiased Cramer-Rao lower-bound, the method appears to be near-optimal for medium to large sample-sizes while having a reasonable complexity through the use of structured computations.

Chapter 5

# Conclusion

This thesis has considered the problem of improving the performance of wireless communication receivers through the use of advanced signal processing with the goal of delivering improved user experience and/or system capacity. The system model considered in this thesis directly reflects the focus on GSM/EDGE and WCDMA/HSPA cellular systems, but methods are directly applicable to general communication systems.

A large portion of the thesis is concerned with approximate detection and decoding and various methods for this are presented. One concept described is that of practical noise/interference whitening by breaking down the covariance matrix to a manageable size. This allows the receiver to reliably model and estimate any colored noise/interference which is not included in a discrete signal model and thus provide flexibility and robustness. Next, the sphere detection and decoding framework is presented and a previously unknown relationship between the QL factorization of the channel transfer matrix and minimum-phase prefiltering is outlined. This connection enables the view of sphere detection as a adaptive variant of traditional reduced-state sequence estimation and thus opening up for interesting possibilities in these two areas. Furthermore, methods based on Generalized Belief Propagation (GBP) for approximate joint detection and decoding in systems with convolutional codes are introduced. In this framework, a generalized Turbo principle is formulated based on the exchange of region beliefs instead of single-variable beliefs as in the ordinary Turbo principle, but

due to time constraints it has not been possible to fully investigate this.

The second part of this thesis is concerned with parameter estimation for the system model considered. A general variational Bayesian EM-algorithm is introduced to the communication community, generalizing the EM-algorithm by propagating parameter distributions instead of point-estimates. By doing this, the parameter uncertainty can be taken into account in the detection/decoding and thus potentially improve performance. Finally, a near-optimal EM-algorithm for performing band-Toeplitz covariance estimation is introduced. This method is generally applicable and directly integrates with the concept of noise and interference whitening.

## 5.1 Suggestions for Further Research

This thesis has presented solutions for approximate detection and decoding, including parameter estimation for such systems. However, in the process of writing this thesis, many new ideas and questions have become apparent and answering these would be interesting. The author believes that several fruitful future research directions exist based on the work presented in this thesis, i.e.

- Investigation of the generalized Turbo principle, also for other setups e.g. decoding of Turbo codes
- General frameworks for loop-correction in GBP
- Unifying the decision-feedback framework with that of sphere detection

Appendix A

# A Low-complexity Joint Synchronization and Detection Algorithm for Single-band DS-CDMA UWB Communications

In EURASIP Journal on Applied Signal Processing, UWB - State of the Art, Issue 3, Pages 462-470, 2005.

This paper is a result of the Master Thesis work done on Ultra-Wide Band (UWB) systems.

EURASIP Journal on Applied Signal Processing 2005:3, 462–470


# A Low-Complexity Joint Synchronization and Detection Algorithm for Single-Band DS-CDMA UWB Communications

**Lars P. B. Christensen**
*Information and Mathematical Modelling, Technical University of Denmark, 2800 Kongens Lyngby, Denmark*
*Email: lc@imm.dtu.dk*

*Received 1 October 2003; Revised 2 June 2004*

The problem of asynchronous direct-sequence code-division multiple-access (DS-CDMA) detection over the ultra-wideband (UWB) multipath channel is considered. A joint synchronization, channel-estimation, and multiuser detection scheme based on the adaptive linear minimum mean square error (LMMSE) receiver is presented and evaluated. Further, a novel nonrecursive least-squares algorithm capable of reducing the complexity of the adaptation in the receiver while preserving the advantages of the recursive least-squares (RLS) algorithm is presented.

**Keywords and phrases:** ultra-wideband, direct-sequence code-division multiple-access, multiuser detection, low-complexity adaptive receivers, synchronization.

## 1. INTRODUCTION

Over the last couple of years, the interest in ultra-wideband (UWB) wireless communications has been growing. Among the reasons for this increased awareness of UWB are the promises of low-power, high-bitrate wireless connections without the need for spectrum allocation, and the approval of the technology by authorities as, for example, the American FCC [1].

UWB signals for wireless communication typically have a bandwidth of several GHz and can be utilized in many ways each presenting the designer with tradeoffs between cost, power, bitrate, range, and the number of users supported. The system considered in this paper is a single-band UWB direct-sequence code-division multiple-access (DS-CDMA) receiver with all signal processing done on the received signal sampled directly from an amplified and filtered antenna signal. This enables the removal of traditional up- and down-converters present in today's narrowband transceivers at the expense of increasing the required sampling rate and thus the complexity of the signal processing. It is therefore of great interest to reduce the complexity of such receivers to make them feasible.

The receiver considered is fully adaptive making it possible to track changes not only in the multipath channel, but also in the received pulse shape. This is desirable in order to maximize performance even under conditions distorting the received pulse shape as discussed in [2], but distortions originating from the electromagnetic propagation environment can also be adaptively compensated for.

Combined LMMSE synchronization and detection for DS-CDMA systems have already been studied (see, e.g., [3, 4, 5, 6, 7]). This paper is a continuation of [8] extended with the synchronization method in [3], but having a low-complexity adaptive algorithm with recursive least-squares (RLS)-speed convergence. Furthermore, this paper uses the channel model presented in [9] instead of the model in [8] as the latter may prove too optimistic for typical office use as a result of the larger dimensions typically present in office environments.

The rest of this paper is organized as follows. Section 2 describes the system model used throughout this paper. In Section 3, the LMMSE receiver is presented as a benchmark of how well the adaptive receiver outlined by Section 4 performs compared to the best possible linear receiver. Synchronization of the receiver is covered in Section 5 and Section 6 presents simulations of the receiver. Section 7 concludes the paper with final remarks.

## 2. SYSTEM MODEL

The receiver considered is the adaptive LMMSE receiver with the system model being capable of supporting $K$ asynchronous users each operating in their respective multipath radio channel. The desired user is, without loss of generality, assumed to be user 1.

### 2.1. Transmitted signal

The pulse shape used for transmission $p(t)$ is of duration $T_{\text{mono}}$ and is assumed normalized to the unit energy. This pulse shape is traditionally called a monocycle in UWB terms and it is typically modeled as the $q$th derivative of a Gaussian pulse [10], which is also the case in this paper. This makes it possible to include the differentiation performed by the antennas and further control the spectrum of the transmitted signal. To include the effect of asynchronous operation between users, the delay $\tau^{(k)}$ is introduced for the $k$th user.

Next, the binary DS spreading code $c^{(k)}(i) \in \{-1, +1\}$, for $i = 1, \ldots, N_c$, is used to separate the different users and provide a processing gain of $N_c$, where $N_c$ indicates the number of coded monocycles transmitted for each bit of information. Finally, the binary information given by $b^{(k)}(j) \in \{-1, +1\}$ is assumed to be a memoryless random source with equal probability of $+1$ and $-1$. The modulation considered is binary phase shift keying (BPSK) and the transmitted signal from the $k$th user can therefore be written as

$$\begin{aligned} s^{(k)}(t) &= \sum_{j=-\infty}^{\infty} b^{(k)}(j)\varphi^{(k)}\left(t - jT_b - \tau^{(k)}\right) \\ &= \sum_{j=-\infty}^{\infty} b^{(k)}(j) \sum_{i=0}^{N_c-1} c^{(k)}(i)p\left(t - jT_b - iT_{\text{mono}} - \tau^{(k)}\right). \end{aligned} \tag{1}$$

The waveform $\varphi^{(k)}(t)$ has duration $T_b = N_c T_{\text{mono}}$ holding $N_c$ monocycles coded by the user's spreading code.

### 2.2. Radio channel

To include the effects of a realistic multipath environment, the radio channel model given in [9] is used. The impulse response of this model for the $k$th user can be written as

$$h^{(k)}(t) = \sum_{l=0}^{L-1} a_l^{(k)} \delta\left(t - lT_{\text{ch}}\right), \tag{2}$$

where $T_{\text{ch}}$ is the temporal spacing between the $L$ multipath components and $\delta(t)$ is the Dirac delta function. The amplitude of the $l$th multipath component is given by $a_l^{(k)}$ and is assumed to be constant over time. Convolving the transmitted signal of the $k$th user given by (1) with its respective impulse response given by (2), the contribution from this user onto the received signal can be written as

$$r^{(k)}(t) = \sum_{l=0}^{L-1} a_l^{(k)} s^{(k)}\left(t - lT_{\text{ch}}\right) \tag{3}$$

and the received signal is therefore

$$\begin{aligned} r(t) &= \sum_{k=1}^{K} r^{(k)}(t) + n(t) \\ &= \sum_{k=1}^{K} \sum_{l=0}^{L-1} a_l^{(k)} s^{(k)}\left(t - lT_{\text{ch}}\right) + n(t) \end{aligned} \tag{4}$$

with $n(t)$ being white Gaussian noise with zero mean and variance $\sigma^2$ leading to the signal-to-noise ratio (SNR) at the receiver being defined as

$$\text{SNR} = \frac{\sum_{l=0}^{L-1} \left| a_l^{(1)} \right|^2}{\sigma^2}. \tag{5}$$

## 3. THE LMMSE RECEIVER

In the receiver an antialiasing filter processes the received signal before it is uniformly sampled and fed directly into a tapped-delay-line filter with the input given by the vector

$$\mathbf{r}(j) = \left[r\left(jT_b\right), r\left(jT_b + T_s\right), \ldots, r\left(jT_b + (N-1)T_s\right)\right]^{\text{T}}, \tag{6}$$

where $N$ is the length of the tapped-delay-line filter with a sample spacing of $T_s$. In order to be able to capture the entire multipath energy spread out by the channel model, the number of filter taps must be at least

$$N_{\text{full}} = \left\lceil \frac{T_b + (L-1)T_{\text{ch}}}{T_s} \right\rceil \tag{7}$$

with the operator $\lceil x \rceil$ returning the smallest integer larger than $x$. However, as the multipath energy tends to decay as a function of the time delay, it may not be cost efficient to capture all the multipath energy from a given bit. A reduction in the filter length is therefore accomplished by setting

$$N = \left\lceil \psi \frac{T_b + (L-1)T_{\text{ch}}}{T_s} \right\rceil, \tag{8}$$

where $0 < \psi \leq 1$ is the filter length reduction compared to the filter that spans the entire multipath energy of a given bit.

The transmitted bits are estimated by hard decision on the output of the filter as

$$\hat{b}^{(1)}(j) = \text{sgn}\left(\mathbf{w}(j)^{\text{T}}\mathbf{r}(j)\right) \tag{9}$$

with $\mathbf{w}(j)$ being the column vector holding the filter coefficients.

In order to evaluate the performance of the LMMSE receiver with perfect knowledge about the channel and user parameters, the contribution from an unmodulated bit can seen to be

$$v^{(k)}(t) = \sum_{l=0}^{L-1} a_l^{(k)} \varphi^{(k)}\left(t - lT_{\text{ch}} - \tau^{(k)}\right) \tag{10}$$

and sampling this signal produces the vector

$$\begin{aligned} &\mathbf{v}^{(k)}(m) \\ &\quad = \left[v^{(k)}\left(mT_b\right), v^{(k)}\left(mT_b + T_s\right), \ldots, v^{(k)}\left(mT_b + (N-1)T_s\right)\right]^{\text{T}}. \end{aligned} \tag{11}$$

Although the expression of (4) includes all bits transmitted, only a finite number of bits, $L_1$ bits before and $L_2$ bits after

the current bit, will contribute energy to $\mathbf{r}(j)$. It is therefore possible to express $\mathbf{r}(j)$ using only the relevant bits as

$$\mathbf{r}(j) = \sum_{k=1}^{K} \sum_{m=-L_1}^{L_2} b^{(k)}(j+m)\mathbf{v}^{(k)}(m) + \mathbf{n}(j) \tag{12}$$

with $\mathbf{n}(j)$ holding the noise samples. The maximum bit offset that contribute energy to $\mathbf{r}(j)$ is therefore

$$L_1 = \left\lceil \frac{(L-1)T_{\text{ch}}}{T_b} \right\rceil \tag{13}$$

as the number of bits in the past influencing the decision is independent of $\psi$. On the other hand, the number of bits after the current bit influencing the decision is

$$L_2 = \left\lceil \psi \frac{(L-1)T_{\text{ch}}}{T_b} \right\rceil. \tag{14}$$

The LMMSE filter coefficients $\mathbf{w}_o$ is given by the Wiener-Hopf solution

$$\mathbf{R}\mathbf{w}_o = \mathbf{p} \Longleftrightarrow \mathbf{w}_o = \mathbf{R}^{-1}\mathbf{p}, \tag{15}$$

where $\mathbf{R}$ is the covariance matrix and $\mathbf{p}$ the cross-correlation vector defined as

$$\begin{aligned} \mathbf{R} &= E[\mathbf{r}(j)\mathbf{r}(j)^{\text{T}}], \\ \mathbf{p} &= E[b^{(1)}(j)\mathbf{r}(j)]. \end{aligned} \tag{16}$$

Applying the expectations of (16) to (12), the covariance matrix can be found to be

$$\mathbf{R} = \sum_{k=1}^{K} \sum_{m=-L_1}^{L_2} \mathbf{v}^{(k)}(m)\mathbf{v}^{(k)}(m)^{\text{T}} + \sigma^2\mathbf{I} \tag{17}$$

with $\mathbf{I}$ being the identity matrix. In a similar way, the cross-correlation vector is found to be

$$\mathbf{p} = \mathbf{v}^{(1)}(0). \tag{18}$$

The output of the filter is

$$\mathbf{w}_o^{\text{T}}\mathbf{r}(j) = \overbrace{\mathbf{w}_o^{\text{T}}\mathbf{v}^{(1)}(0)}^{\text{Desired}} + \overbrace{e_{\text{ISI}}(j) + e_{\text{MAI}}(j)}^{\text{Interference}} + \overbrace{e_n(j)}^{\text{Noise}}, \tag{19}$$

where $e_{\text{ISI}}(j)$, $e_{\text{MAI}}(j)$, and $e_n(j)$ are the contributions at the output from intersymbol interference (ISI), multiple-access interference (MAI), and noise, respectively. Both $e_{\text{ISI}}(j)$ and $e_{\text{MAI}}(j)$ are approximately Gaussian as shown in [11] and $e_n(j)$ is Gaussian as the filter is linear. The BER of the LMMSE receiver may therefore be approximated by

$$\text{BER}_{\text{LMMSE}} = \frac{1}{2}\,\text{erfc}\left(\sqrt{\frac{\left|\mathbf{w}_o^{\text{T}}\mathbf{v}^{(1)}(0)\right|^2}{2\left(\sigma_{\text{ISI}}^2 + \sigma_{\text{MAI}}^2 + \sigma^2\right)}}\right) \tag{20}$$

with $\sigma^2$ being the noise variance and

$$\begin{aligned} \sigma_{\text{ISI}}^2 &= \sum_{m\neq 0} \left|\mathbf{w}_o^{\text{T}}\mathbf{v}^{(1)}(m)\right|^2, \\ \sigma_{\text{MAI}}^2 &= \sum_{k=2}^{K} \sum_{m=-L_1}^{L_2} \left|\mathbf{w}_o^{\text{T}}\mathbf{v}^{(k)}(m)\right|^2. \end{aligned} \tag{21}$$

## 4. THE ADAPTIVE LMMSE RECEIVER

Instead of implementing the LMMSE receiver by performing matrix inversion, the filter coefficients can be obtained by adaptation of the filter using an appropriate training sequence. The normalized least mean square (NLMS) and RLS algorithms are presented here only to give a better understanding of the nonrecursive formulation of the RLS algorithm presented later in this section. For all algorithms, the filter coefficients are initialized to the zero vector, that is, $\mathbf{w}(0) = \mathbf{0}$.

### 4.1. The NLMS algorithm

The NLMS update can be written as [12]

$$\mathbf{w}(j+1) = \mathbf{w}(j) + \kappa(j)\mathbf{r}(j)e(j), \tag{22}$$

where $e(j)$ is the a posteriori error given by

$$e(j) = b^{(1)}(j) - \mathbf{w}(j)^{\text{T}}\mathbf{r}(j). \tag{23}$$

The variable $\kappa(j)$ controls the effective step-size and is found as

$$\kappa(j) = \frac{\mu}{a + \mathbf{r}(j)^{\text{T}}\mathbf{r}(j)}, \quad a \ll E[\mathbf{r}(j)^{\text{T}}\mathbf{r}(j)] \tag{24}$$

with $\mu$ being the step-size bound to the interval $0 < \mu < 2$ by stability. The constant $a$ is introduced to reduce the impact of gradient noise when $\mathbf{r}(j)^{\text{T}}\mathbf{r}(j)$ attains a small value. The choice of the step-size parameter $\mu$ is a tradeoff between convergence speed, and thus the needed number of training bits, and the residual error resulting in an increased BER compared to the value of (20).

### 4.2. The RLS algorithm

The RLS update can be written as [12]

$$\mathbf{w}(j) = \mathbf{w}(j-1) + \overbrace{\mathbf{\Phi}^{-1}(j)\mathbf{r}(j)}^{\mathbf{k}(j)}\varepsilon(j) \tag{25}$$

with $\mathbf{\Phi}(j)$ being the sample covariance matrix defined by

$$\mathbf{\Phi}(j) = \frac{1}{j}\sum_{i=1}^{j} \mathbf{r}(i)\mathbf{r}(i)^{\text{T}} \tag{26}$$

and

$$\varepsilon(j) = b^{(1)}(j) - \mathbf{w}(j-1)^{\text{T}}\mathbf{r}(j) \tag{27}$$

being the a priori error. In order to reduce the complexity of the RLS update to approximately $O(4N^2)$ per bit, the following recursion is used:

$$\mathbf{k}(j) = \frac{\mathbf{\Phi}^{-1}(j-1)\mathbf{r}(j)}{1+\mathbf{r}(j)^{\mathrm{T}}\mathbf{\Phi}^{-1}(j-1)\mathbf{r}(j)}, \tag{28}$$

$$\mathbf{\Phi}^{-1}(j) = \mathbf{\Phi}^{-1}(j-1) - \mathbf{k}(j)\mathbf{r}(j)^{\mathrm{T}}\mathbf{\Phi}^{-1}(j-1). \tag{29}$$

Initialization of the inverse covariance matrix is done as

$$\mathbf{\Phi}^{-1}(0) = \frac{\delta}{E[\mathbf{r}(j)^{\mathrm{T}}\mathbf{r}(j)]}\mathbf{I} \simeq \frac{\delta}{\mathbf{r}(0)^{\mathrm{T}}\mathbf{r}(0)}\mathbf{I}, \tag{30}$$

where $\delta$ is a regularization parameter. A value of $\delta \ll 1$ will cause a high degree of regularization whereas $\delta \gg 1$ will introduce little regularization. The choice of $\delta$ is therefore a tradeoff between reducing the noise and not constraining the adaptation.

### 4.3. The nonrecursive least-squares algorithm

The nonrecursive least-squares (NLS) algorithm will now be derived from the RLS update. Let the vector $\gamma(j)$ be defined as

$$\gamma(j) = \mathbf{\Phi}^{-1}(j-1)\mathbf{r}(j) \tag{31}$$

and rewrite (29) as

$$\mathbf{\Phi}^{-1}(j) = \mathbf{\Phi}^{-1}(j-1) - \frac{\gamma(j)\gamma(j)^{\mathrm{T}}}{\delta(j)} \tag{32}$$

with the scalar $\delta(j)$ being defined as

$$\begin{aligned}\delta(j) &= 1+\mathbf{r}(j)^{\mathrm{T}}\mathbf{\Phi}^{-1}(j-1)\mathbf{r}(j)\\ &= 1+\mathbf{r}(j)^{\mathrm{T}}\gamma(j).\end{aligned} \tag{33}$$

Using these definitions, it is possible to rewrite the RLS update as

$$\mathbf{w}(j) = \mathbf{w}(j-1) + \gamma(j)\frac{\varepsilon(j)}{\delta(j)}. \tag{34}$$

The idea is now to rewrite (31) using (32) and expand the expression all the way back to the first iteration, that is, $j = 1$ resulting in

$$\gamma(j) = \mathbf{\Phi}^{-1}(0)\mathbf{r}(j) + \sum_{i=1}^{j-1}\frac{1}{\delta(i)}\gamma(i)\gamma(i)^{\mathrm{T}}\mathbf{r}(j). \tag{35}$$

However, instead of using the usual recursive formulation of (35), having a complexity of $O(4N^2)$, the nonrecursive version as directly outlined by (35) has a complexity of $O(3(j-1)N)$ at the $j$th iteration. This formulation of the RLS algorithm takes advantage of the fact that at the $j$th iteration, the rank of the sample covariance matrix is only $j-1$, if the initialization matrix is not considered, and only $j-1$ inner products are therefore needed to get $\gamma(j)$.

The ratio $G(j)$ between the complexity of the RLS and NLS algorithms at the $j$th iteration is approximately

$$G(j) \simeq \frac{4N^2}{3(j-1)N} = \frac{4N}{3(j-1)} \tag{36}$$

and the NLS algorithm is therefore beneficial if convergence is reached in less than approximately $4N/3$ iterations. Further, the complexity reduction averaged over the performed iterations is $2G(N_{\mathrm{ite}})$ with $N_{\mathrm{ite}}$ being the number of iterations performed as the algorithm has a lower complexity in the first iterations. Therefore, using the overall complexity as a measure, the NLS algorithm is beneficial if convergence is reached within approximately $8N/3$ iterations.

In many signal processing problems, the rank of the covariance matrix is full or close to being full, leading to slow convergence of the RLS algorithm. If this is the case, the nonrecursive implementation is not preferable over the usual recursive implementation. However, when the rank is low compared to the dimension of the covariance matrix, a considerable reduction of complexity is possible as a result of the higher speed of convergence. An example of such a problem is the adaptive receiver considered in this paper.

### 4.4. The windowed NLS algorithm

Another interesting aspect of the nonrecursive formulation is the possibility to limit the number of summations per iteration as

$$\gamma(j) = \mathbf{\Phi}^{-1}(0)\mathbf{r}(j) + \sum_{i=j-D}^{j-1}\frac{1}{\delta(i)}\gamma(i)\gamma(i)^{\mathrm{T}}\mathbf{r}(j), \quad i>0, \tag{37}$$

where $D$ is the number of terms included, resulting in a complexity of $O(3DN)$ per iteration when disregarding the initialization matrix. The algorithm now performs a minimization of the squared error over a sliding rectangular window of size $D$, that is,

$$\arg\min_{\mathbf{w}(j)}\left(\sum_{i=j-D-1}^{j}|\varepsilon(i)|^2\right), \quad i>0. \tag{38}$$

The algorithm is therefore termed the windowed NLS (WNLS) algorithm. Window functions other than the rectangular one specified here can of course also be used if desired. The algorithm can be considered a kind of a generalization of the NLMS and RLS algorithms as $D = 0$ equals the NLMS algorithm and $D = j-1$ equals the RLS algorithm. Values of $D$ in between these two extremes provide algorithms with convergence speed scaling with $D$ as the algorithm estimates the sample covariance matrix over the window. It should also be noticed that when $j \leq D+1$, the WNLS algorithm is equivalent to the NLS algorithm.

## 5. SYNCHRONIZATION OF THE ADAPTIVE LMMSE RECEIVER

The task of synchronizing the receiver with the transmitter and staying synchronized over time is an often-overlooked

topic compared to modulation and demodulation. However, as this is absolutely crucial to the performance of the system, a method of synchronizing the adaptive LMMSE receiver is presented here based on the same principles as used in [3].

The type of synchronization considered is the initial synchronization including both bit and frame synchronization over the UWB multipath channel in [9]. However, the problem of tracking changes between the transmitter and the receiver is not considered. It is therefore assumed that the clocks of the receiver and transmitter are the same except for an unknown offset and that the channel is stationary.

### 5.1. Bit synchronization

Firstly, bit synchronization can be established by taking advantage of the adaptive nature of the receiver. If at first the AWGN channel is observed, it can be noted that if the receiver is not synchronized to the transmitter, extending the filter length by one bit length can capture all energy from a desired bit. The adaptive algorithm will therefore automatically suppress coefficients outside of the correct bit interval and bit synchronization is therefore automatically achieved, but this comes at the expense of increasing the filter length to twice its original size. Increasing the filter length by a bit length in the UWB multipath channel will, in a similar way as in the AWGN channel, ensure that at least the same energy is captured as if the systems were synchronous. It is then possible to estimate the timing offset between the transmitter and receiver by observing the converged filter coefficients and use this to correct the timing in the receiver [7]. In this manner, the receiver will be able to take full advantage of the increased filter length to capture a larger part of the multipath energy, but this correction is not included in this paper.

The increase in filter length may be modeled by a larger value of $\psi$ given by

$$\psi' = \psi + \psi_b, \tag{39}$$

where $\psi$ determines the filter length of the fully synchronous system and $\psi_b$ represents the increase needed to accommodate a full bit length and is given by

$$\psi_b = \frac{T_{\text{mono}} N_c}{T_s N_{\text{full}}} = \frac{N_c}{N_c + (L-1)T_{\text{ch}}/T_{\text{mono}}}. \tag{40}$$

The AWGN channel therefore requires $\psi_b = 1$ as argued earlier and in the case of the UWB multipath channel, the value of $\psi_b$ will typically be much less than unity and the increase in complexity will therefore be small. This is a direct consequence of the fact that the energy spread in the UWB channel is typically much larger than the bit period.

### 5.2. Frame synchronization

In order for the receiver to lock onto the transmitted information, the bits are arranged into a frame consisting of $N_f$ bits. In the beginning of the frame, a known length $N_t$ maximal-length sequence is inserted acting as a synchronization burst to make the adaptation possible. The remaining $N_d = N_f - N_t$ bits of the frame are the information bits. However, as the receiver has no knowledge of when to look for the synchronization sequence, this ambiguity can be modeled by placing the start of the synchronization burst at a position $N_s$ unknown to the receiver.

To acquire correct synchronization, the receiver will now have to estimate $N_s$. This is done by searching all possible positions of the synchronization burst and select the estimate $\hat{N}_s$ that leads to the smallest mean square error (MSE) averaged over the performed iterations, that is,

$$\arg\min_{\hat{N}_s} \sum_{j=1}^{N_t} \left| b^{(1)}(j) - \mathbf{w}(j-1)^{\mathrm{T}} \mathbf{r}\left(j + \hat{N}_s\right) \right|^2. \tag{41}$$

The receiver now uses the converged coefficients at the estimated position to detect the transmitted bits. Since the current bit influences the observation window as long as $-L_2 \le e_s \le L_1$, it is not required that the synchronization error $e_s = N_s - \hat{N}_s$ be zero in order to correctly detect a bit. Still, having $e_s = 0$ maximizes the received energy and thus makes it desirable to minimize $|e_s|$.

## 6. SIMULATION AND DISCUSSION

A number of simulations have been performed to assess the performance of the described UWB receiver in the multipath channel specified in [9].

The used monocycle is the 7th derivative of a Gaussian pulse with a pulse width $T_{\text{mono}} = 0.67$ nanosecond, as the spectrum of this pulse propagating in free space is a good match for the FCC regulations [1] giving a bandwidth in the order of 3 GHz [13]. The number of samples per monocycle was set to 13 yielding $T_s = 51.3$ picoseconds in order to provide good rejection of aliasing at half the sample rate. It may however be possible to reduce this high sampling rate by taking advantage of the aliasing in the form of sub-Nyquist sampling [8].

The system simulated consists of $K$ sample-asynchronous users each using a length $N_c = 15$ large-set Kasami spreading code, making it possible for up to approximately 15 users to simultaneously use the system. The users do not need to have knowledge about the spreading codes used in the system, as the receiver requires only the training sequence to adapt. All users are assumed received at the same power level.

The channel model employs a tap spacing of $T_{\text{ch}} = 2$ nanoseconds with the total number of taps being $L = 100$ [9]. This results in the number of filter coefficients being $N_{\text{full}} = 4056$ if the entire energy spread in the channel is to be covered. The channel impulse response is fixed during adaptation and BER measurements, but to help average out the stochastic nature of the channel model, simulations are averaged over 10 different channels. The reason for using only 10 different channels is that it is computationally

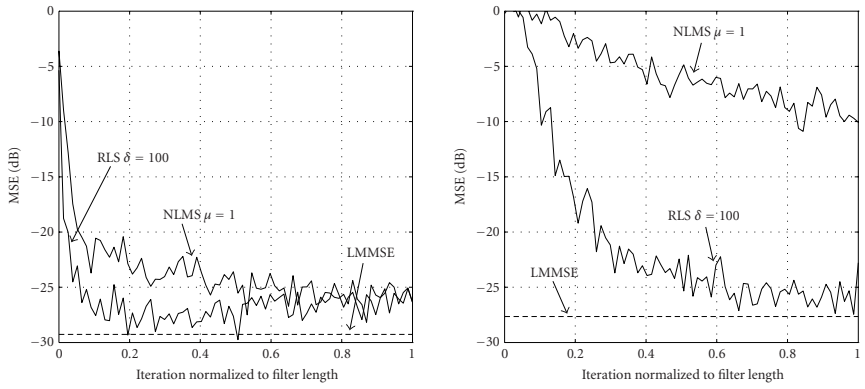


FIGURE 1: Convergence of the receiver ($N_c = 15$, $\psi = 1$, SNR = 20 dB). (a) $K = 1$ and (b) $K = 15$.

intractable to average out the entire channel and that this number of channels drawn from the model produces results being within ±0.5 dB of the results obtained by performing the much larger number of simulations needed to average out the channel distribution.

For NLMS, a step-size of $\mu = 1$ was selected, as a smaller step-size will produce unacceptable slow convergence. In the case of RLS, the value $\delta = 100$ was chosen to minimize the effect of regularization as it is of higher importance not to constrain the adaptation when many users are active in the UWB multipath channel.

For a more in-depth description of the effects of these adaptation parameters on the performance of the system in both the AWGN and UWB multipath channel, the interested reader is referred to [13].

### 6.1. Convergence

The convergence behavior of the receiver is important in order to determine the number of training bits necessary and verify that the filter coefficients converge to the LMMSE solution.

Observing the convergence plotted in Figure 1, it should be noted how the addition of users makes the receiver converge more slowly as the dimension of the problem scales with the number of users. In the case of 15 users using the NLMS adaptation, the speed of convergence becomes very slow and does not reach convergence within the simulated iterations. The RLS algorithm manages to converge much faster as a result of its knowledge of the estimated inverse covariance matrix, but increasing the number of users also impacts it.

In Figure 2a, the convergence of the WNLS algorithm is plotted showing how the performance scales from NLMS to RLS when increasing the window length, as its knowledge of the estimated inverse covariance matrix grows with the window length.

### 6.2. BER simulations

A series of Monte Carlo simulations have been performed to estimate the BER performance of the receiver under the assumption that the receiver has knowledge of the timing parameter $\tau^{(1)}$. The number of iterations performed is kept fixed at $N_{\text{ite}} = N_{\text{full}}$ and a total of 100 bit errors must occur before a BER value is accepted.

From Figure 3 it can be seen that under both light- and full-load conditions of 1 and 15 users, respectively, the RLS algorithm is capable of providing reasonably good performance even in the case of restricting the filter length to approximately $\psi = 0.2$. In the case of only a single user, the RLS algorithm comes very close to the LMMSE receiver, but it is not quite capable of reaching the bound when the load is increased to 15 users. The NLMS algorithm has been left out, as its general performance is unsatisfying [13], which is also clear from the slow convergence depicted in Figure 1.

### 6.3. Synchronization

By inserting the needed parameters in (40), the filter length can be seen to increase by $\psi_b = 0.048$ in order to let the filter span an extra bit length. Focusing on the case of $\psi = 0.2$ this results in $\psi' = 0.248$ leading to $L_1 = 20$ and $L_2 = 5$. The BER performance of the receiver with this extended filter length is plotted in Figure 3 under the assumption of being synchronized with the desired user.

The performance of the joint synchronization and detection is shown in Figure 4 assuming $N_d = 500$. Further,

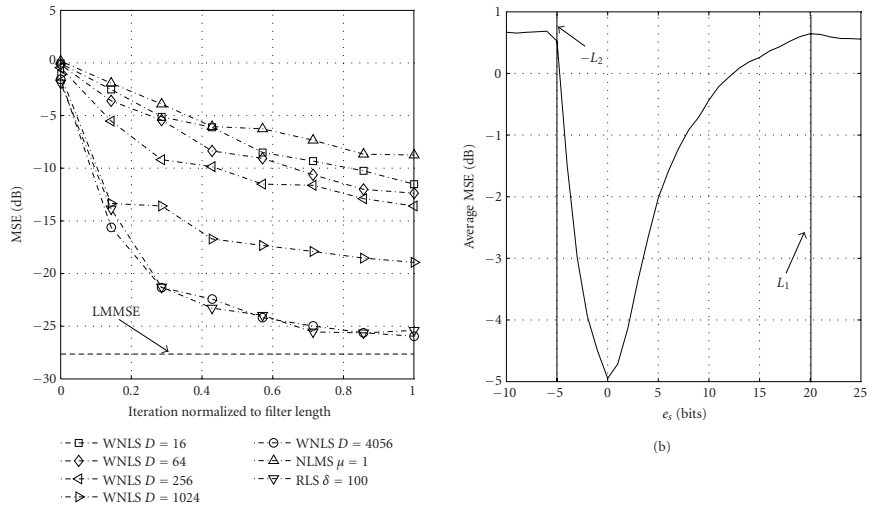


FIGURE 2: Convergence of the WNLS algorithm and the average MSE as a function of synchronization error ($N_c = 15$, $\psi' = 0.248$, $\delta = 100$). (a) $K = 15$, SNR = 20 dB and (b) $K = 1$, SNR = 10 dB, $N_{\text{ite}} = N_t = 127$.

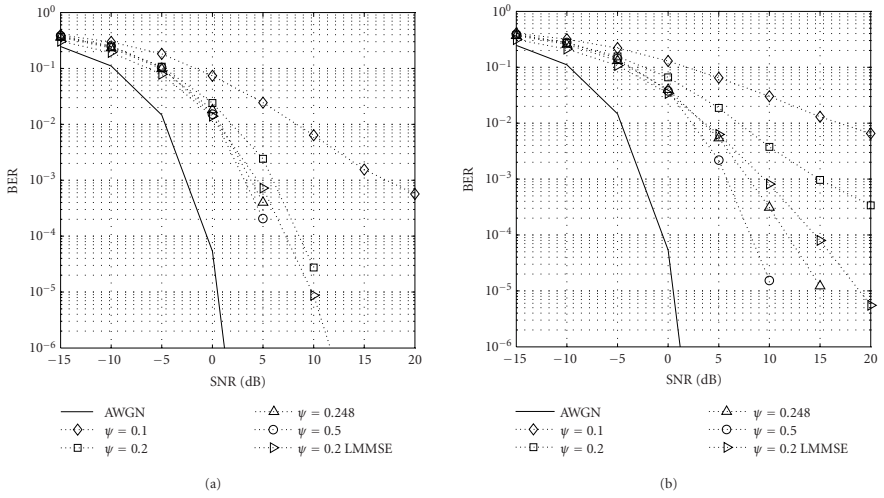


FIGURE 3: The BER in the UWB multipath channel when the receiver is synchronized to the desired user ($N_c = 15$, $N_{\text{ite}} = N_{\text{full}} = 4056$, RLS $\delta = 100$). (a) $K = 1$ and (b) $K = 15$.

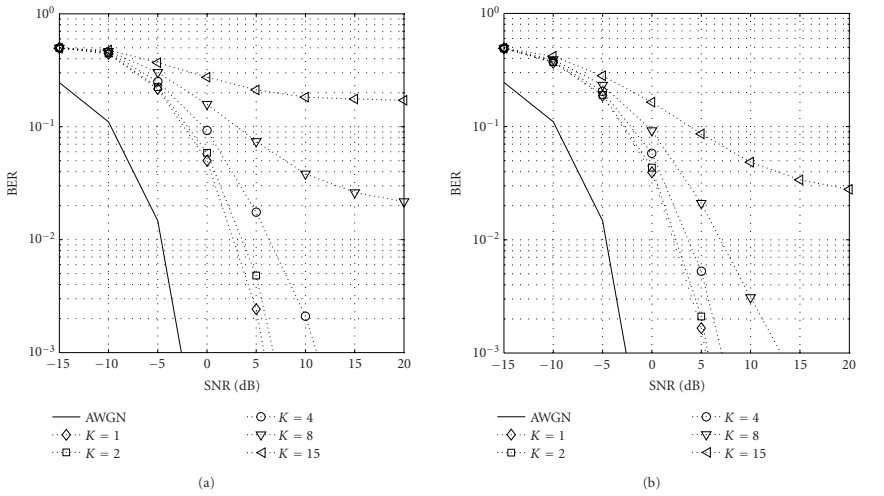


FIGURE 4: Performance of the presented joint synchronization and detection scheme using the NLS algorithm ($N_c = 15$, $\psi' = 0.248$, $\delta = 100$). (a) $N_{\text{ite}} = N_t = 127$ and (b) $N_{\text{ite}} = N_t = 255$.

Figure 2b plots the average MSE as a function of the synchronization error showing how on average the synchronization error is minimized by (41). However, the synchronization error may be nonzero and the performance of the receiver therefore degrades, as the captured energy becomes less. This, along with the fact that in the two cases shown only $N_{\text{ite}} = 127$ and $N_{\text{ite}} = 255$ iterations are performed, explains why the BER in Figure 4 degrades compared to that of Figure 3, especially when more users are added. This performance degradation is the price paid by using this low-complexity type of joint synchronization and detection. However, the achieved performance is the same as could be reached by using the RLS algorithm, but in the example where $N_{\text{ite}} = 127$, the NLS algorithm lowers the complexity by a factor of $G(N_{\text{ite}}) \simeq 10$ resulting in approximately 20 times the overall complexity reduction.

## 7. CONCLUSION

A method for performing joint synchronization, channel estimation, and multiuser detection for single-band DS-CDMA UWB communications has been presented based on the principles in [3, 8]. Simulations of the receiver show good results in the UWB multipath channel in [9] using RLS adaptation, but the complexity of the RLS adaptation is very high. To help alleviate this problem, a novel algorithm termed the WNLS algorithm is derived, potentially lowering the computational complexity while preserving the performance of the RLS algorithm.

## ACKNOWLEDGMENTS

The author would like to thank the anonymous reviewers for pointing out that the synchronization scheme was already in existence, as the author was unaware of this fact. Further, the author wishes to thank Thomas Fabricius, Spectronic Denmark, and Associate Professor Jan Larsen, Technical University of Denmark, for their various fruitful discussions. In addition, the author greatly appreciates the careful proofreading by Pedro Højen-Sørensen, Nokia Denmark, and Ole Nørklit, Nokia Denmark.

## REFERENCES

[1] Federal Communications Commission (FCC), "Revision of part 15 of the commission's rules regarding ultra-wideband transmission systems," First Report and Order, ET Docket 98-153, FCC 02-48; Adopted: February 2002; Released: April 2002.

[2] A. Muqaibel, A. Safaai-Jazi, B. Woerner, and S. Riad, "UWB channel impulse response characterization using deconvolution techniques," in *Proc. 45th IEEE Midwest Symposium on Circuits and Systems (MWSCAS '02)*, vol. 3, pp. 605–608, Tulsa, Okla, USA, August 2002.

[3] U. Madhow, "Adaptive interference suppression for joint acquisition and demodulation of direct-sequence CDMA signals," in *Proc. IEEE Military Communications Conference (MILCOM '95)*, vol. 3, pp. 1200–1204, San Diego, Calif, USA, November 1995.

[4] U. Madhow, "MMSE interference suppression for timing acquisition and demodulation in direct-sequence CDMA systems," *IEEE Trans. Commun.*, vol. 46, no. 8, pp. 1065–1075, 1998.

[5] R. Smith and S. Miller, "Acquisition performance of an adaptive receiver for DS-CDMA," *IEEE Trans. Commun.*, vol. 47, no. 9, pp. 1416–1424, 1999.
[6] M. Latva-aho, J. Lilleberg, J. Iinatti, and M. Juntti, "CDMA downlink code acquisition performance in frequency-selective fading channels," in *Proc. 9th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '98)*, vol. 3, pp. 1476–1480, Boston, Mass, USA, September 1998.
[7] M. El-Tarhuni and A. Sheikh, "Performance analysis for an adaptive filter code-tracking technique in direct-sequence spread-spectrum systems," *IEEE Trans. Commun.*, vol. 46, no. 8, pp. 1058–1064, 1998.
[8] Q. Li and L. A. Rusch, "Multiuser detection for DS-CDMA UWB in the home environment," *IEEE J. Select. Areas Commun.*, vol. 20, no. 9, pp. 1701–1711, 2002.
[9] D. Cassioli, M. Z. Win, and A. F. Molisch, "The ultra-wide bandwidth indoor channel: from statistical model to simulations," *IEEE J. Select. Areas Commun.*, vol. 20, no. 6, pp. 1247–1257, 2002.
[10] M. Z. Win and R. A. Scholtz, "Impulse radio: how it works," *IEEE Commun. Lett.*, vol. 2, no. 2, pp. 36–38, 1998.
[11] H. V. Poor and S. Verdu, "Probability of error in MMSE multiuser detection," *IEEE Trans. Inform. Theory*, vol. 43, no. 3, pp. 858–871, 1997.
[12] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 1996.
[13] L. P. B. Christensen, "Signal processing for ultra-wideband systems," M.S. thesis, Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, May 2003, http://www.imm.dtu.dk/~lc.

**Lars P. B. Christensen** was born in Copenhagen, Denmark, in November 1978. He received the M.S.E.E. degree from the Technical University of Denmark in May 2003, where he is currently working towards the Ph.D. degree. His current research interests are in the areas of digital communications and statistical signal processing.

Appendix B

# Minimum Symbol Error Rate for SIMO Channels with Markov Noise

In IEEE 6th Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 236-240, 5-8 June 2005.

# MINIMUM SYMBOL ERROR RATE DETECTION IN SINGLE-INPUT MULTIPLE-OUTPUT CHANNELS WITH MARKOV NOISE

*Lars P. B. Christensen*

Technical University of Denmark, email: lc@imm.dtu.dk

**ABSTRACT**

Minimum symbol error rate detection in Single-Input Multiple-Output(SIMO) channels with Markov noise is presented. The special case of zero-mean Gauss-Markov noise is examined closer as it only requires knowledge of the second-order moments. In this special case, it is shown that optimal detection can be achieved by a Multiple-Input Multiple-Output(MIMO) whitening filter followed by a traditional BCJR algorithm. The Gauss-Markov noise model provides a reasonable approximation for co-channel interference, making it an interesting single-user detector for many multi-user communication systems where interference from other transmitters has a limiting effect.

## 1. INTRODUCTION

Interference from other users is a limiting factor in many real-life communication systems. The optimal solution is to jointly detect the desired and interfering users, but this has a complexity scaling exponentially with the number of interferers. Approaching the optimal performance with lower complexity is therefore of great interest and much work has been done within this field of research.

The idea followed in this work is to only detect the desired user and model the rest as noise. The solution to this problem of single-user minimum symbol error rate detection in channels with memory, today known simply as the BCJR algorithm after its inventors, is derived in [1] for memoryless noise. In [2] and [3], a Markov model of the noise is assumed and the optimal symbol-by-symbol and sequence detectors are derived, but only for Single-Input Single-Output(SISO) channels.

However, optimal representation of communication signals may require multiple observations per symbol. This is the case, if the signal has a bandwidth beyond the Nyquist frequency giving rise to a cyclostationary signal after digital sampling [4]. Another example is, if multiple antennas are available in the receiver. The resulting redundancy in the signal should therefore be exploited to better reject the interfering signal.

This work was supported by Nokia Denmark.

The presented framework is easily extended to include MIMO channels and/or multi-user detection by extending the discrete state-space, but this is outside the scope of this paper. Group detection of some streams/users and letting the remaining be approximated as Markov noise is also an option. Furthermore, the probabilistic nature of the detector makes it a good match for iterative decoding and parameter estimation schemes.

Section 2 presents the signal model and Section 3 derives the optimal symbol-by-symbol detector for SIMO channels with memoryless noise in order to establish the formalism that Section 4 extends to Markov noise. Next, the special case of SIMO channels with Gauss-Markov noise is looked into in Section 5. Section 6 contains simulation results of the presented detector using a GSM physical layer to highlight the benefit of the detector under the influence of co-channel interference.

## 2. SIGNAL MODEL

The received signal is assumed to have $N_d$ receive dimensions and the received signal in the $d^{th}$ dimension is modeled as

$$r_d(t) = \sum_{n=-\infty}^{\infty} h_d(t - nT)\, s_n + \epsilon_d(t) \qquad (1)$$

with $T$ being the symbol period and $\epsilon_d(t)$ being the noise process of the $d^{th}$ receive dimension. The $n^{th}$ complex symbol $s_n$ belongs to the constellation set $\Omega$ and the number of constellation points is $|\Omega|$. Further, it is assumed that the constellation points have unit average power, i.e. $\Omega = \{+1, -1\}$ for binary modulation leading to $|\Omega| = 2$. The overall time-invariant impulse response $h_d(\tau)$ in the $d^{th}$ receive dimension includes pulse shaping in the transmitter, radio propagation channel and receive filtering. Furthermore, it is assumed to be zero for $\tau < 0$ and having finite temporal length $LT$.

In the detector, the received signal is sampled once every $T$ for each dimension and the samples from all dimensions at the $n^{th}$ symbol is then put into a stacked column vector

notation as

$$\mathbf{r}(nT) = [r_0(nT), r_1(nT), ..., r_{N_d-1}(nT)]^T \quad (2)$$

where $(\cdot)^T$ indicates matrix transpose. Samples starting from the $n_1$'th symbol and ending at the $n_2$'th symbols is represented by

$$\mathbf{r}_{n_1}^{n_2} = [\mathbf{r}(n_1T)^T, \mathbf{r}((n_1+1)T)^T, \ldots, \mathbf{r}(n_2T)^T]^T \quad (3)$$

Assuming $s_n = 0$ for $n < 0$ and that $N_s - L + 1$ symbols have been transmitted, the available received signal of length $N_sT$ is therefore $\mathbf{r}_0^{N_s-1}$.

## 3. OPTIMAL DETECTION IN SIMO CHANNELS WITH MEMORYLESS NOISE

The optimal symbol-by-symbol detector chooses the symbol having the highest posterior probability, that is

$$\hat{s}_n = \arg\max_{s_n} P\left(s_n \mid \mathbf{r}_0^{N_s-1}\right) \quad (4)$$

As a result of the finite constellation size $|\Omega|$ and the channel having a memory of $L-1$ symbols, the task of finding the posterior distribution may be formulated as an inference problem on a Hidden Markov Model(HMM) with $|\Omega|^{L-1}$ states. The state vector of the $j^{th}$ state at the $n^{th}$ symbol determine the symbols interfering with the desired user in $\mathbf{r}_n^n$ assuming the $j^{th}$ combination of the interfering symbols. The state vector is therefore defined as

$$\boldsymbol{\sigma}_n^j = [\hat{s}_{n-L+1}^j, \hat{s}_{n-L+2}^j, ..., \hat{s}_{n-1}^j] \quad (5)$$

with $\hat{s}_{n-1}^j$ being the value of the $n-1$'th symbol in the $j^{th}$ state. Further define $\Omega_0$ as the set of all states and $\Omega(s_{n-1})$ as the set of states where $\hat{s}_{n-1}$ takes on a specific value from $\Omega$. This makes it possible to rewrite the posterior probability as

$$\begin{aligned} P\left(s_{n-1} \mid \mathbf{r}_0^{N_s-1}\right) &= \frac{P\left(s_{n-1}, \mathbf{r}_0^{N_s-1}\right)}{P\left(\mathbf{r}_0^{N_s-1}\right)} \\ &= \frac{\sum_{\boldsymbol{\sigma}_n^j \in \Omega(s_{n-1})} P\left(\mathbf{r}_0^{N_s-1}, \boldsymbol{\sigma}_n^j\right)}{\sum_{\boldsymbol{\sigma}_n^j \in \Omega_0} P\left(\mathbf{r}_0^{N_s-1}, \boldsymbol{\sigma}_n^j\right)} \end{aligned} \quad (6)$$

The result of Equation (6) is, that the desired posterior can be extracted from the joint probability $P\left(\mathbf{r}_0^{N_s-1}, \boldsymbol{\sigma}_n^j\right)$ by marginalization. Finding the joint probability is done by splitting it into a forward variable $\alpha_n^j$ and a backward variable $\beta_n^j$ as

$$\begin{aligned} P\left(\mathbf{r}_0^{N_s-1}, \boldsymbol{\sigma}_n^j\right) &= P\left(\mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j\right) P\left(\mathbf{r}_n^{N_s-1} \mid \mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j\right) \\ &= \underbrace{P\left(\mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j\right)}_{\alpha_n^j} \underbrace{P\left(\mathbf{r}_n^{N_s-1} \mid \boldsymbol{\sigma}_n^j\right)}_{\beta_n^j} \end{aligned} \quad (7)$$

The conditioning on $\mathbf{r}_0^{n-1}$ can be dropped as the state vector contains all information about the past. However, this is only true when the noise is memoryless as assumed here.

To find the forward and backward variables, let the apriori probability of a state transition be $P\left(\boldsymbol{\sigma}_n^j \mid \boldsymbol{\sigma}_{n-1}^{j'}\right)$ and the set of states where $P\left(\boldsymbol{\sigma}_n^j \mid \boldsymbol{\sigma}_{n-1}^{j'}\right) \neq 0$ as $\Omega_n$. This makes it possible to find the forward variable recursively as

$$\begin{aligned} \alpha_n^j = P\left(\mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j\right) &= \sum_{\boldsymbol{\sigma}_{n-1}^{j'} \in \Omega_n} P\left(\mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right) \\ &= \sum_{\boldsymbol{\sigma}_{n-1}^{j'} \in \Omega_n} \underbrace{P\left(\mathbf{r}_0^{n-2}, \boldsymbol{\sigma}_{n-1}^{j'}\right)}_{\alpha_{n-1}^{j'}} P\left(\boldsymbol{\sigma}_n^j \mid \boldsymbol{\sigma}_{n-1}^{j'}\right) \\ &\quad P\left(\mathbf{r}_{n-1}^{n-1} \mid \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right) \end{aligned} \quad (8)$$

and the backward variable as

$$\begin{aligned} \beta_n^j = P\left(\mathbf{r}_n^{N_s-1} \mid \boldsymbol{\sigma}_n^j\right) &= \sum_{\boldsymbol{\sigma}_{n+1}^{j'} \in \Omega_{n+1}} P\left(\mathbf{r}_n^{N_s-1}, \boldsymbol{\sigma}_{n+1}^{j'} \mid \boldsymbol{\sigma}_n^j\right) \\ &= \sum_{\boldsymbol{\sigma}_{n+1}^{j'} \in \Omega_{n+1}} \underbrace{P\left(\mathbf{r}_{n+1}^{N_s-1} \mid \boldsymbol{\sigma}_{n+1}^{j'}\right)}_{\beta_{n+1}^{j'}} P\left(\boldsymbol{\sigma}_{n+1}^{j'} \mid \boldsymbol{\sigma}_n^j\right) \\ &\quad P\left(\mathbf{r}_n^n \mid \boldsymbol{\sigma}_{n+1}^{j'}, \boldsymbol{\sigma}_n^j\right) \end{aligned} \quad (9)$$

In the special case where the noise $\epsilon_d(t)$ is zero-mean Additive White Gaussian Noise(AWGN) with variance $\sigma^2$ per dimension, the observation probability is conditionally Gaussian and is given by

$$\begin{aligned} &-2ln\left(P\left(\mathbf{r}_{n-1}^{n-1} \mid \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)\right) + Z \\ &\quad = \frac{1}{\sigma^2}\left\|\mathbf{r}_{n-1}^{n-1} - \hat{\mathbf{r}}_{n-1}^{n-1}\left(\mathbf{h}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)\right\|^2 \end{aligned} \quad (10)$$

with the normalization constant $Z = N_d ln\left(2\pi\sigma^2\right)$. The function $\hat{\mathbf{r}}_{n-1}^{n-1}\left(\mathbf{h}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)$ reconstructs the desired signal from the sampled impulse response $\mathbf{h}$ and the overall state. In the case of AWGN, the complexity of the algorithm is $\mathcal{O}\left(N_d|\Omega|^L\right)$ operations per symbol as $\mathcal{O}\left(N_d|\Omega|\right)$ operations must be performed in each of the $|\Omega|^{L-1}$ states.

## 4. OPTIMAL DETECTION IN SIMO CHANNELS WITH MARKOV NOISE

The noise process $\epsilon_d(t)$ is now assumed to be Markov with a finite temporal memory $N_mT$ and possibly coupled across

the receive dimensions. As before, it is desirable to model the desired signal by a HMM, but now the noise is independently modeled by a Markov model and the memory needed to correctly capture the state of the system is now $L-1+N_m$ as the memories add up. The number of states is now $|\Omega|^{L-1+N_m}$ and the new state vector is therefore defined as

$$\boldsymbol{\sigma}_n^j = [\hat{s}_{n-L-N_m+1}^j, \hat{s}_{n-L-N_m+2}^j, ..., \hat{s}_{n-1}^j] \tag{11}$$

As in Equation (7), the joint probability is found by splitting it into a forward variable and backward variable by

$$\begin{aligned} P\left(\mathbf{r}_0^{N_s-1}, \boldsymbol{\sigma}_n^j\right) &= P\left(\mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j\right) P\left(\mathbf{r}_n^{N_s-1} \mid \mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j\right) \\ &= \underbrace{P\left(\mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j\right)}_{\alpha_n^j} \underbrace{P\left(\mathbf{r}_n^{N_s-1} \mid \mathbf{r}_{n-N_m}^{n-1}, \boldsymbol{\sigma}_n^j\right)}_{\beta_n^j} \end{aligned} \tag{12}$$

However, the conditioning on $\mathbf{r}_0^{n-1}$ can only be reduced to $\mathbf{r}_{n-N_m}^{n-1}$ as a result of the finite memory Markov noise and the fact that only the desired signal is described by $\boldsymbol{\sigma}_n^j$.

As before the forward variable $\alpha_n^j$ and the backward variable $\beta_n^j$ may be found recursively. The forward variable is found by

$$\begin{aligned} \alpha_n^j &= P\left(\mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j\right) = \sum_{\boldsymbol{\sigma}_{n-1}^{j'} \in \Omega_n} P\left(\mathbf{r}_0^{n-1}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right) \\ &= \sum_{\boldsymbol{\sigma}_{n-1}^{j'} \in \Omega_n} \underbrace{P\left(\mathbf{r}_0^{n-2}, \boldsymbol{\sigma}_{n-1}^{j'}\right)}_{\alpha_{n-1}^{j'}} P\left(\boldsymbol{\sigma}_n^j \mid \boldsymbol{\sigma}_{n-1}^{j'}\right) \\ &\qquad P\left(\mathbf{r}_{n-1}^{n-1} \mid \mathbf{r}_{n-1-N_m}^{n-2}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right) \end{aligned} \tag{13}$$

and the backward variable by

$$\begin{aligned} \beta_n^j &= P\left(\mathbf{r}_n^{N_s-1} \mid \mathbf{r}_{n-N_m}^{n-1}, \boldsymbol{\sigma}_n^j\right) \\ &= \sum_{\boldsymbol{\sigma}_{n+1}^{j'} \in \Omega_{n+1}} P\left(\mathbf{r}_n^{N_s-1}, \boldsymbol{\sigma}_{n+1}^{j'} \mid \mathbf{r}_{n-N_m}^{n-1}, \boldsymbol{\sigma}_n^j\right) \\ &= \sum_{\boldsymbol{\sigma}_{n+1}^{j'} \in \Omega_{n+1}} \underbrace{P\left(\mathbf{r}_{n+1}^{N_s-1} \mid \mathbf{r}_{n+1-N_m}^{n}, \boldsymbol{\sigma}_{n+1}^{j'}\right)}_{\beta_{n+1}^{j'}} \\ &\qquad P\left(\boldsymbol{\sigma}_{n+1}^{j'} \mid \boldsymbol{\sigma}_n^j\right) P\left(\mathbf{r}_n^n \mid \mathbf{r}_{n-N_m}^{n-1}, \boldsymbol{\sigma}_{n+1}^{j'}, \boldsymbol{\sigma}_n^j\right) \end{aligned} \tag{14}$$

The observation probability is now conditioned on previous received samples and a convenient way of avoiding this is by rewriting using Bayes' rule as

$$\begin{aligned} &P\left(\mathbf{r}_{n-1}^{n-1} \mid \mathbf{r}_{n-1-N_m}^{n-2}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right) \\ &\qquad = \frac{P\left(\mathbf{r}_{n-1-N_m}^{n-1} \mid \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)}{P\left(\mathbf{r}_{n-1-N_m}^{n-2} \mid \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)} \end{aligned} \tag{15}$$

which can be evaluated directly. The complexity of the algorithm is $\mathcal{O}\left(|\Omega|^{L+N_m}\right)$ per symbol excluding the complexity involved in evaluating Equation (15).

## 5. OPTIMAL DETECTION IN SIMO CHANNELS WITH GAUSS-MARKOV NOISE

The special case where the noise is not only Markov, but also zero-mean Gaussian is described in this section. This model fits communication systems well and is interesting as only the covariance matrix of the noise must be known in order to evaluate Equation (15). Any Gauss-Markov process having non-zero mean is also described by this section, as it may be transformed into having zero-mean by subtracting the mean.

Let the noise samples be stacked in the same manner in $\boldsymbol{\epsilon}_{n_1}^{n_2}$ as in $\mathbf{r}_{n_1}^{n_2}$ for the received signal. To evaluate the top of Equation (15), the covariance

$$\boldsymbol{\Sigma} = E\left[\boldsymbol{\epsilon}_{n-N_m}^{n}\left(\boldsymbol{\epsilon}_{n-N_m}^{n}\right)^H\right] \tag{16}$$

is required where $(\cdot)^H$ indicates matrix transpose and complex conjugation. The lower part requires the covariance

$$\breve{\boldsymbol{\Sigma}} = E\left[\boldsymbol{\epsilon}_{n-N_m}^{n-1}\left(\boldsymbol{\epsilon}_{n-N_m}^{n-1}\right)^H\right] \tag{17}$$

However, the last covariance is included in the first as can be seen by block partioning the first as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \breve{\boldsymbol{\Sigma}} & \mathbf{B} \\ \mathbf{B}^H & \mathbf{A} \end{bmatrix} \tag{18}$$

Defining

$$\mathbf{W} = \boldsymbol{\Sigma}^{-1} - \begin{bmatrix} \breve{\boldsymbol{\Sigma}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{19}$$

makes it possible to rewrite Equation (15) as

$$\begin{aligned} &-2ln\left(\frac{P\left(\mathbf{r}_{n-1-N_m}^{n-1} \mid \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)}{P\left(\mathbf{r}_{n-1-N_m}^{n-2} \mid \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)}\right) + Z = \\ &\hat{\boldsymbol{\epsilon}}_{n-1-N_m}^{n-1}\left(\mathbf{h}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)^H \mathbf{W} \hat{\boldsymbol{\epsilon}}_{n-1-N_m}^{n-1}\left(\mathbf{h}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right) \end{aligned} \tag{20}$$

with the normalization constant being

$$Z = ln\left|\mathbf{A} - \mathbf{B}^H \breve{\mathbf{\Sigma}}^{-1}\mathbf{B}\right| + N_d ln\left(2\pi\right) \tag{21}$$

where $|\cdot|$ indicates the matrix determinant and the estimated noise is defined as

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}_{n-1-N_m}^{n-1}\left(\mathbf{h}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right) \\ = \mathbf{r}_{n-1-N_m}^{n-1} - \hat{\mathbf{r}}_{n-1-N_m}^{n-1}\left(\mathbf{h}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)\end{aligned} \tag{22}$$

However, the rank of $\mathbf{W}$ is not full. It is shown in Section 8 that $rank\left(\mathbf{W}\right) = N_d$, thus making it possible to express $\mathbf{W}$ as a sum of $N_d$ vector outer products

$$\begin{aligned}\mathbf{W} &= \sum_{d=0}^{N_d-1} \mathbf{f}_d \mathbf{f}_d^H \\ &= \underbrace{\left[\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_{N_d-1}\right]}_{\mathbf{F}} \underbrace{\left[\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_{N_d-1}\right]^H}_{\mathbf{F}^H}\end{aligned} \tag{23}$$

Whitening of the noise process can therefore be achieved by the MIMO whitening filter $\mathbf{F}^H$. The whitened received signal and reconstructed signal are defined as

$$\begin{aligned}\tilde{\mathbf{r}}_{n-1}^{n-1} &= \mathbf{F}^H \mathbf{r}_{n-1-N_m}^{n-1} \\ \tilde{\hat{\mathbf{r}}}_{n-1}^{n-1}\left(\mathbf{h}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right) &= \mathbf{F}^H \hat{\mathbf{r}}_{n-1-N_m}^{n-1}\left(\mathbf{h}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)\end{aligned} \tag{24}$$

making it possible to rewrite Equation (20) as

$$\begin{aligned}-2ln\left(\frac{P\left(\mathbf{r}_{n-1-N_m}^{n-1} \mid \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)}{P\left(\mathbf{r}_{n-1-N_m}^{n-2} \mid \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)}\right) + Z \\ = \left\|\tilde{\mathbf{r}}_{n-1}^{n-1} - \tilde{\hat{\mathbf{r}}}_{n-1}^{n-1}\left(\mathbf{h}, \boldsymbol{\sigma}_n^j, \boldsymbol{\sigma}_{n-1}^{j'}\right)\right\|^2\end{aligned} \tag{25}$$

This is simply the AWGN version of the receiver as the signal has now been whitened by $\mathbf{F}^H$, but the channel length is now $L + N_m$ instead of $L$. The complexity of the receiver using the whitening filter is therefore $\mathcal{O}\left(N_d \left|\Omega\right|^{L+N_m}\right)$ operations per symbol excluding finding and applying the whitening filter. As expected, the results of [2] and [3] are recovered if $N_d = 1$ is inserted.

## 6. NUMERICAL SIMULATIONS

In this section, a number of simulations using a GSM physical layer is presented using the zero-mean Gauss-Markov model to approximate Co-Channel Interference(CCI). All interferers are fully synchonized in time with the desired user. The used modulation is Gaussian Minimum Shift Keying(GMSK) with $BT = 0.3$ in accordance with the GSM specifications. However, as this modulation is non-linear, it is linearized to fit the model of Equation (1) by the Laurent approximation [5]. The channel model used for all simulations is the Typical Urban(TU) model given by the GSM specifications using a mobile speed of 50 km/h at a carrier frequency of 945 MHz. All simulations are performed using perfect knowledge of the linearized impulse response of the desired user and covariance matrix of the linearized interference.

In the detector the received signal is oversampled by a factor of $N_{sps}$ relative to the symbolrate. There are several reasons for having $N_{sps} > 1$ in a GSM receiver, one being that the bandwidth of the signal exceeds the Nyquist frequency. After sampling, the received signal is derotated to remove the rotation in the GMSK modulation [5]. Next, the received signal is split into a real and imaginary part as this improves the rejection of interferers with real constellation points, such as derotated GMSK, by exploiting redundancy in the interfering signal [6]. The number of real-valued receive dimensions for the receiver is therefore $N_d = 2N_{sps}$.

As a measure of the level of interference, the Carrier-to-Interference Ratio($CIR$) is defined as the average received power for the desired user divided by the average received interference power. Further, AWGN is added to account for any thermal noise with the Signal-to-Noise Ratio($SNR$) being defined as the average received signal power from the desired user divided by the noise variance.

For comparison, the performance of IQ-LMMSE[6] and BCJR detection without whitening is shown using perfect parameter estimates. Like the Gauss-Markov approximation, the IQ-LMMSE relies on second-order moments to suppress interference and therefore seems to be a reasonable comparison.

In Figure 1 the Bit Error Rate(BER) of a GSM user is plotted under the influence of one GMSK co-channel interferer. It can be seen that a gain can be achieved by increasing the oversampling from $N_{sps} = 1$ to $N_{sps} = 2$ and that the IQ-LMMSE is outperformed, especially at higher values of $CIR$. Higher values of $N_{sps}$ are not included as only marginal gains compared to $N_{sps} = 2$ are achievable for this scenario. The value of $L = 7$ is selected in order to capture all energy spread out in time by the channel model and $N_m = 3$ was selected as higher values seem to provide little improvement.

In Figure 2 another interferer is added with the second interferer being 10dB weaker than the first. The performance is clearly impacted by the additional interferer, but a significant gain is still achieved over IQ-LMMSE and by using $N_{sps} = 2$ compared to $N_{sps} = 1$.

## 7. CONCLUSION

Minimum symbol error rate detection in SIMO channels with Markov noise has been derived as a generalization of

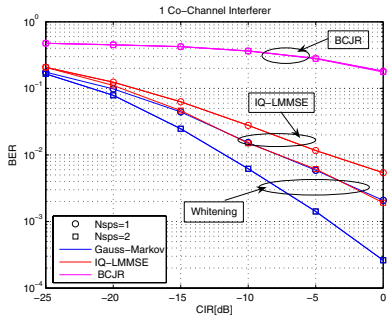


**Fig. 1**. BER plot of 1 CCI on TU50 channels with $L = 7$, $N_m = 3$, $SNR = 40dB$.

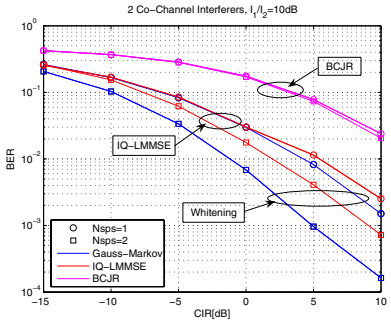


**Fig. 2**. BER plot of 2 CCIs on TU50 channels with $L = 7$, $N_m = 3$, $SNR = 40dB$.

the work presented in [2]. In the special case of Gauss-Markov noise, it has been shown that optimal detection can be achieved by a MIMO whitening filter followed by a traditional BCJR algorithm. Numerical simulations have shown that the presented detector can improve the performance compared to the IQ-LMMSE and the detector in [2] when multiple observations per symbol are available in interference-limited scenarios.

## 8. PROOF OF RANK

In this section it is shown that $rank\,(\mathbf{W}) = N_d$ with $\mathbf{W}$ defined by Equations (18)-(19). This is most easily done by using the inversion lemma for block partitioned matrices to yield

$$\begin{aligned}\mathbf{W} &= \mathbf{\Sigma}^{-1} - \begin{bmatrix} \breve{\mathbf{\Sigma}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\mathbf{B}}\mathbf{S}\tilde{\mathbf{B}}^H & -\tilde{\mathbf{B}}\mathbf{S} \\ -\mathbf{S}\tilde{\mathbf{B}}^H & \mathbf{S} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{B}} \\ -\mathbf{I} \end{bmatrix} \mathbf{S} \begin{bmatrix} \tilde{\mathbf{B}} \\ -\mathbf{I} \end{bmatrix}^H \end{aligned} \tag{26}$$

with

$$\begin{aligned}\tilde{\mathbf{B}} &= \breve{\mathbf{\Sigma}}^{-1}\mathbf{B} \\ \mathbf{S} &= \left(\mathbf{A} - \mathbf{B}^H\breve{\mathbf{\Sigma}}^{-1}\mathbf{B}\right)^{-1}\end{aligned} \tag{27}$$

where $\mathbf{I}$ indicates the identity matrix. As the Schur complement $\mathbf{S}$ is positive definite, it may be Cholesky factorized as $\mathbf{S} = \mathbf{C}\mathbf{C}^H$ making it possible to rewrite Equation (26) as

$$\mathbf{W} = \underbrace{\begin{bmatrix} \tilde{\mathbf{B}} \\ -\mathbf{I} \end{bmatrix} \mathbf{C}}_{\mathbf{F}} \underbrace{\mathbf{C}^H \begin{bmatrix} \tilde{\mathbf{B}} \\ -\mathbf{I} \end{bmatrix}^H}_{\mathbf{F}^H} \tag{28}$$

The result in Equation (28) directly shows $rank\,(\mathbf{W}) = N_d$ and thereby completes the proof.

## 9. REFERENCES

[1] F. Jelinek L. R. Bahl, J. Cocke and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. on Info. Theory*, pp. 284–287, March 1974.

[2] A. Kavcic, "Soft-output detector for channels with intersymbol interference and markov noise memory," in *IEEE Global Telecommunications Conference*, 1999, vol. 1B, pp. 728–732.

[3] A. Kavcic and J. M. F. Moura, "The viterbi algorithm and markov noise memory," *IEEE Trans. on Info. Theory*, vol. 46, pp. 291–301, January 2000.

[4] J. G. Proakis, *Digital Communications*, McGraw-Hill, 3rd edition, 1995.

[5] P. Laurent, "Exact and approximate construction of digital phase modulations by superposition of amplitude modulated pulses (amp)," *IEEE Trans. on Communications*, vol. 34, no. 2, pp. 150–160, February 1986.

[6] R. Schober W. H. Gerstacker and A. Lampe, "Receivers with widely linear processing for frequency-selective channels," *IEEE Trans. on Communications*, vol. 51, no. 9, pp. 1512–1523, September 2003.

Appendix C

# On Data and Parameter Estimation Using the VBEM-algorithm for Block-fading Frequency-selective MIMO Channels

In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) vol. 4, pp. 465-468, 2006.

# ON DATA AND PARAMETER ESTIMATION USING THE VARIATIONAL BAYESIAN EM-ALGORITHM FOR BLOCK-FADING FREQUENCY-SELECTIVE MIMO CHANNELS

*Lars P. B. Christensen, Jan Larsen*

Informatics and Mathematical Modelling, Technical University of Denmark
Email: {lc, jl}@imm.dtu.dk

**ABSTRACT**

A general Variational Bayesian framework for iterative data and parameter estimation for coherent detection is introduced as a generalization of the EM-algorithm. Explicit solutions are given for MIMO channel estimation with Gaussian prior and noise covariance estimation with inverse-Wishart prior. Simulation of a GSM-like system provides empirical proof that the VBEM-algorithm is able to provide better performance than the EM-algorithm. However, if the posterior distribution is highly peaked, the VBEM-algorithm approaches the EM-algorithm and the gain disappears. The potential gain is therefore greatest in systems with a small amount of observations compared to the number of parameters to be estimated.

## 1. INTRODUCTION

The focus of this paper is on improved iterative data and parameter estimation for coherent detection in block-fading frequency-selective MIMO channels. Much work has been done within this field and many variants of the EM-algorithm have been applied to communication systems, see for example [1, 2, 3]. However, previous estimators have all provided point-estimates of the parameters, not distributions as offered by the full Bayesian approach. On the other hand, Bayesian estimators average over the distribution of the unknown variables or parameters to provide improved inference about the system. Previously, a so-called Bayesian EM (BEM)-algorithm was introduced for communication systems [2, 3]. However, the BEM-algorithm provides a Maximum A Posteriori (MAP) point-estimate and is therefore not a true Bayesian estimator.

The contribution of this paper is to introduce the Variational Bayesian EM (VBEM)-algorithm, already used extensively in the machine-learning community, to the communications society. Explicitly, the contribution is to formulate an iterative data, channel and noise covariance estimator based on the VBEM-algorithm. By simulations it is shown, that the performance of a communication system can be improved over that based on the EM-algorithm when there is significant uncertainty in the parameter estimates.

The first author would like to thank Nokia Denmark for funding

## 2. SYSTEM MODEL

We will consider the uncoded linear $M \times N$ MIMO system

$$\mathbf{y}_i = \mathbf{H}\mathbf{x}_i + \mathbf{n}_i \tag{1}$$

where $\mathbf{H} \in \mathbb{C}^{M\times N}$ is the channel matrix and $\mathbf{x}_i \in \Omega^{N\times 1}$ is the vector of transmitted symbols at time index $i$, each belonging to the complex-valued alphabet $\Omega$. The received signal vector $\mathbf{y}_i \in \mathbb{C}^{M\times 1}$ holds the observations at time $i$ and the additive noise $\mathbf{n}_i \in \mathbb{C}^{M\times 1}$ is assumed to be circular zero-mean Gaussian with covariance $\mathbf{\Sigma} \triangleq E\left[\mathbf{n}_i\mathbf{n}_i^H\right]$ and $E\left[\mathbf{n}_i\mathbf{n}_i^T\right] = \mathbf{0}$. The generalization of the estimation framework to Gauss-Markov noise is straightforward [4].

The frequency-selective channel is assumed to have a temporal length of $L$ symbols. Let $N_t$ and $N_r$ denote the number of transmitters and receivers respectively leading to $N = LN_t$ and $M = N_r$. For channel estimation, it is desirable to rewrite the channel matrix into a vector notation as

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{h} + \mathbf{n}_i \tag{2}$$

with $\mathbf{h} \triangleq vec\,(\mathbf{H})$ where $vec\,(\cdot)$ is the column stacking operator. The $k$'th row of the symbol matrix $\mathbf{X}_i \in \mathbb{C}^{N_r\times LN_tN_r}$ is found by upsampling $\mathbf{x}_i^T$ by $N_r$ and shifting it right by $k-1$ positions producing a Toeplitz structure. The two representations are equivalent and we can use the best suited depending on conditions.

Assuming data is sent in frames of $N_f$ symbols per transmitter, the collection of all transmitted symbols and observations is given by

$$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_f}\}, \quad \mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{N_e}\} \tag{3}$$

where $N_e = N_f + L - 1$ due to the convolutive multipath channel.

## 3. MAXIMUM LIKELIHOOD ESTIMATION

In this section, a quick outline of Maximum Likelihood (ML) estimation using the EM-algorithm is presented as the VBEM-algorithm is a generalization of the EM-algorithm. The framework is in a general form and is carried over to the formulation of the VBEM-algorithm. For further details, see [5, 6].

The idea behind the EM-algorithm is to consider the observations $\mathbf{y}$ being incomplete data as the underlying hidden variables $\mathbf{x}$ are unknown. This problem is overcome by considering the hidden variables as being random variables and averaging over their distribution. By this philosophy we can write the complete-data log-likelihood of the parameter set $\boldsymbol{\theta}$

$$E: \quad Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j-1)}\right) \triangleq \langle ln\left[p\left(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta}\right)\right]\rangle_{p\left(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{(j-1)}\right)} \quad (4)$$

where $\boldsymbol{\theta}^{(j-1)}$ is the parameter set from the previous iteration and $\langle\cdot\rangle_{p(\cdot)}$ indicates averaging w.r.t. the distribution in the subscript. Carrying out the above averaging is often termed the E-step. Next, in the so-called M-step we maximize w.r.t. $\boldsymbol{\theta}$, i.e.

$$M: \quad \boldsymbol{\theta}^{(j)} \triangleq \arg\max_{\boldsymbol{\theta}} \quad Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j-1)}\right) \quad (5)$$

We now have an iterative algorithm, which can be shown to converge to a local maximum in $p\left(\mathbf{y}, \boldsymbol{\theta}\right)$. However, the fact that the algorithm "only" converges to a local maximum makes initialization of the algorithm crucial, as it may otherwise converge to an incorrect maximum.

In terms of the system model from Section 2, the observations are the received samples $\mathbf{y}$, the hidden variables correspond to the transmitted symbols $\mathbf{x}$ and the parameter set is $\boldsymbol{\theta} = \{\mathbf{h}, \boldsymbol{\Sigma}\}$. In the E-step, the posterior distribution of the transmitted symbols $p\left(\mathbf{x} \mid \mathbf{y}, \mathbf{h}^{(j)}, \boldsymbol{\Sigma}^{(j)}\right)$ is found by the well-known BCJR algorithm using forward-backward recursions, see e.g. [4]. The M-step finds the joint ML channel and covariance estimate, but this produces non-linear systems of equations that in the general case appear to have no closed-form solution. A solution is to find the individual ML estimates and possibly iterate between them in the M-step. The individual solutions are easily found to be the Weighted Least-Squares estimator and the sample covariance for the channel and covariance estimate respectively, both averaged over the posterior of the symbols. This common result is not reproduced here, but is given by $\mathbf{h}_{MAP}^{(j)}$ in (13) and $\mathbf{S}^{(j)}$ in (14) by replacing the parameter distribution with a delta-function in the ML point estimate, i.e. $q_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right) = \delta\left(\boldsymbol{\theta} - \boldsymbol{\theta}_{ML}\right)$ and $\boldsymbol{\Sigma}_1^{-1} = \mathbf{0}$.

## 4. BAYESIAN ESTIMATION

In a truly Bayesian framework, all unknown variables and parameters are treated as random variables with some distribution that can be integrated out. The marginal likelihood of the model is therefore found by integrating out the uncertainty as

$$p\left(\mathbf{y}\right) = \int p\left(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}\right) d\mathbf{x} d\boldsymbol{\theta} \quad (6)$$

However, for interesting models the integration is likely to be intractable as it involves multi-dimensional integrals over complicated expressions. Instead, we lower-bound the marginal log-likelihood by Jensen's inequality as

$$\begin{aligned} ln\left[p\left(\mathbf{y}\right)\right] &= ln\left[\int q\left(\mathbf{x}, \boldsymbol{\theta}\right) \frac{p\left(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}\right)}{q\left(\mathbf{x}, \boldsymbol{\theta}\right)} d\mathbf{x} d\boldsymbol{\theta}\right] \\ &\geq \int q\left(\mathbf{x}, \boldsymbol{\theta}\right) ln\left[\frac{p\left(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}\right)}{q\left(\mathbf{x}, \boldsymbol{\theta}\right)}\right] d\mathbf{x} d\boldsymbol{\theta} \end{aligned} \quad (7)$$

where $q\left(\mathbf{x}, \boldsymbol{\theta}\right)$ is a free distribution used to approximate the posterior $p\left(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}\right)$. Maximizing the lower-bound w.r.t. the free distribution $q\left(\mathbf{x}, \boldsymbol{\theta}\right)$ yields the exact posterior, which was what we started out with, and is therefore of no interest. Constraining the free distribution to factorize between the hidden variables and the parameters by requiring

$$q\left(\mathbf{x}, \boldsymbol{\theta}\right) = q_{\mathbf{x}}\left(\mathbf{x}\right) q_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right) \quad (8)$$

provides the intriguing solution that we can optimize the free distributions individually and iterate between them to maximize the lower-bound. This is done by the alternating between the VBE-step and the VBM-step given by

$$\begin{aligned} VBE: \quad & q_{\mathbf{x}}^{(j)}\left(\mathbf{x}\right) \propto e^{\langle ln[p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})]\rangle_{q_{\boldsymbol{\theta}}^{(j-1)}(\boldsymbol{\theta})}} \\ VBM: \quad & q_{\boldsymbol{\theta}}^{(j)}\left(\boldsymbol{\theta}\right) \propto p\left(\boldsymbol{\theta}\right) e^{\langle ln[p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})]\rangle_{q_{\mathbf{x}}^{(j)}(\mathbf{x})}} \end{aligned} \quad (9)$$

where $p\left(\boldsymbol{\theta}\right)$ is a parameter prior. Due to the factorization, global convergence can not be guaranteed, but it can be shown to converge to a local maximum in $p\left(\mathbf{y}\right)$. From (9) we see that the VBEM-algorithm is similar to the EM-algorithm, but the distinction between hidden variables and unknown parameters has vanished as the VBE- and VBM-steps are both averaging over posterior distributions. For more details on Bayesian estimation and the VBEM-algorithm, see [6, 7].

Returning to the system model of Section 2 we now have

$$q\left(\mathbf{x}, \mathbf{h}, \boldsymbol{\Sigma}\right) = q_{\mathbf{x}}\left(\mathbf{x}\right) q_{\mathbf{h}}\left(\mathbf{h}\right) q_{\boldsymbol{\Sigma}}\left(\boldsymbol{\Sigma}\right) \quad (10)$$

where the free distribution is further assumed to factorize between the channel and noise covariance posterior. This approach is equivalent to the individual maximization described in Section 3 for the M-step. The above facorization can be seen to yield the updates

$$\begin{aligned} q_{\mathbf{x}}^{(j)}\left(\mathbf{x}\right) &\propto e^{\langle ln[p(\mathbf{y}, \mathbf{x}|\mathbf{h}, \boldsymbol{\Sigma})]\rangle_{q_{\mathbf{h}}^{(j-1)}(\mathbf{h}) q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})}} \\ q_{\mathbf{h}}^{(j)}\left(\mathbf{h}\right) &\propto p\left(\mathbf{h}\right) e^{\langle ln[p(\mathbf{y}, \mathbf{x}|\mathbf{h}, \boldsymbol{\Sigma})]\rangle_{q_{\mathbf{x}}^{(j)}(\mathbf{x}) q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})}} \\ q_{\boldsymbol{\Sigma}}^{(j)}\left(\boldsymbol{\Sigma}\right) &\propto p\left(\boldsymbol{\Sigma}\right) e^{\langle ln[p(\mathbf{y}, \mathbf{x}|\mathbf{h}, \boldsymbol{\Sigma})]\rangle_{q_{\mathbf{x}}^{(j)}(\mathbf{x}) q_{\mathbf{h}}^{(j)}(\mathbf{h})}} \end{aligned} \quad (11)$$

To simplify the updates, the parameter priors should be conjugate meaning that the posterior is of the same type as the prior. For the channel estimate, the conjugate prior is $\mathbf{h} \sim \mathcal{CN}\left(\mathbf{h}_1, \boldsymbol{\Sigma}_1\right)$ and for the covariance, it is the inverse-Wishart distribution [8].

For the channel estimate, using (2)-(3) and the fact that the noise and prior is Gaussian, we get

$$
\begin{aligned}
&- ln\left[q_{\mathbf{h}}^{(j)}(\mathbf{h})\right] + Z_1 = (\mathbf{h}-\mathbf{h}_1)^H \boldsymbol{\Sigma}_1^{-1}(\mathbf{h}-\mathbf{h}_1) \\
&\quad + \sum_{i=1}^{N_e} \left\langle (\mathbf{y}_i - \mathbf{X}_i\mathbf{h})^H \left\langle \boldsymbol{\Sigma}^{-1} \right\rangle_{q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})} (\mathbf{y}_i - \mathbf{X}_i\mathbf{h}) \right\rangle_{q_{\mathbf{x}}^{(j)}(\mathbf{x})}
\end{aligned} \tag{12}
$$

with $Z$ indicating a normalization constant. Due to the choice of a conjugate prior, the posterior is Gaussian and given by $q_{\mathbf{h}}^{(j)}(\mathbf{h}) \sim \mathcal{CN}\left(\mathbf{h}_{MAP}^{(j)}, \boldsymbol{\Sigma}_{\mathbf{h}}^{(j)}\right)$ with covariance and mean

$$
\begin{aligned}
&\boldsymbol{\Sigma}_{\mathbf{h}}^{(j)} = \left( \sum_{i=1}^{N_e} \left\langle \mathbf{X}_i^H \left\langle \boldsymbol{\Sigma}^{-1} \right\rangle_{q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})} \mathbf{X}_i \right\rangle_{q_{\mathbf{x}}^{(j)}(\mathbf{x})} + \boldsymbol{\Sigma}_1^{-1} \right)^{-1} \\
&\mathbf{h}_{MAP}^{(j)} = \\
&\quad \boldsymbol{\Sigma}_{\mathbf{h}}^{(j)} \left( \sum_{i=1}^{N_e} \left\langle \mathbf{X}_i^H \right\rangle_{q_{\mathbf{x}}^{(j)}(\mathbf{x})} \left\langle \boldsymbol{\Sigma}^{-1} \right\rangle_{q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})} \mathbf{y}_i + \boldsymbol{\Sigma}_1^{-1}\mathbf{h}_1 \right)
\end{aligned} \tag{13}
$$

The distribution of the noise covariance is

$$
\begin{aligned}
&- ln\left[q_{\boldsymbol{\Sigma}}^{(j)}(\boldsymbol{\Sigma})\right] + ln\left[p(\boldsymbol{\Sigma})\right] - N_e ln|\boldsymbol{\Sigma}| + Z_2 \\
&= \sum_{i=1}^{N_e} \left\langle (\mathbf{y}_i - \mathbf{X}_i\mathbf{h})^H \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{h}) \right\rangle_{q_{\mathbf{x}}^{(j)}(\mathbf{x}) q_{\mathbf{h}}^{(j)}(\mathbf{h})} \\
&= tr\left\{ \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N_e} \left\langle (\mathbf{y}_i - \mathbf{X}_i\mathbf{h})(\mathbf{y}_i - \mathbf{X}_i\mathbf{h})^H \right\rangle_{q_{\mathbf{x}}^{(j)}(\mathbf{x}) q_{\mathbf{h}}^{(j)}(\mathbf{h})} \right\} \\
&= tr\left\{ \boldsymbol{\Sigma}^{-1} \mathbf{S}^{(j)} \right\}
\end{aligned} \tag{14}
$$

where $\mathbf{S}^{(j)}$ is the sample covariance averaged over the posteriors. It can be shown [8], that for the inverse-Wishart prior $\boldsymbol{\Sigma} \sim \mathcal{CW}^{-1}(\nu, \boldsymbol{\Sigma}_2)$, we get

$$
\left\langle \boldsymbol{\Sigma}^{-1} \right\rangle_{q_{\boldsymbol{\Sigma}}^{(j)}(\boldsymbol{\Sigma})} = (N_e + \nu)\left(\mathbf{S}^{(j)} + \boldsymbol{\Sigma}_2\right)^{-1} \tag{15}
$$

which for the noninformative prior $\boldsymbol{\Sigma} \sim \mathcal{CW}^{-1}(0, \mathbf{0})$ is equivalent to the ML covariance estimate. The conjugate priors can therefore be interpreted as inserting virtual observations into the estimation.

The VBE-step is similar to the traditional BCJR algorithm, only now we average over the posterior distribution of the parameters. The required state transition probabilities $\gamma(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\theta})$ are therefore of the form

$$
\begin{aligned}
&- ln\left[\gamma(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\theta})\right] + Z_3 \\
&\quad = \left\langle (\mathbf{y}_i - \mathbf{X}_i\mathbf{h})^H \left\langle \boldsymbol{\Sigma}^{-1} \right\rangle_{q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})} (\mathbf{y}_i - \mathbf{X}_i\mathbf{h}) \right\rangle_{q_{\mathbf{h}}^{(j-1)}(\mathbf{h})} \\
&\quad = Z_4 - 2Re\left\{ \mathbf{y}_i^H \left\langle \boldsymbol{\Sigma}^{-1} \right\rangle_{q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})} \mathbf{X}_i \left\langle \mathbf{h} \right\rangle_{q_{\mathbf{h}}^{(j-1)}(\mathbf{h})} \right\} \\
&\qquad + tr\left\{ \left\langle \mathbf{h}\mathbf{h}^H \right\rangle_{q_{\mathbf{h}}^{(j-1)}(\mathbf{h})} \mathbf{X}_i^H \left\langle \boldsymbol{\Sigma}^{-1} \right\rangle_{q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})} \mathbf{X}_i \right\}
\end{aligned} \tag{16}
$$

As the posterior distribution of the channel estimate is Gaussian, we have

$$
\begin{aligned}
\left\langle \mathbf{h} \right\rangle_{q_{\mathbf{h}}^{(j)}(\mathbf{h})} &= \mathbf{h}_{MAP}^{(j)} \\
\boldsymbol{\Sigma}_m^{(j)} \triangleq \left\langle \mathbf{h}\mathbf{h}^H \right\rangle_{q_{\mathbf{h}}^{(j)}(\mathbf{h})} &= \mathbf{h}_{MAP}^{(j)} \left(\mathbf{h}_{MAP}^{(j)}\right)^H + \boldsymbol{\Sigma}_{\mathbf{h}}^{(j)}
\end{aligned} \tag{17}
$$

Inserting this into (16), we get

$$
\begin{aligned}
&- ln\left[\gamma(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\theta})\right] + Z_3 - Z_4 \\
&\quad = -2Re\left\{ \mathbf{y}_i^H \left\langle \boldsymbol{\Sigma}^{-1} \right\rangle_{q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})} \mathbf{X}_i \mathbf{h}_{MAP}^{(j-1)} \right\} \\
&\qquad + tr\left\{ \boldsymbol{\Sigma}_m^{(j-1)} \mathbf{X}_i^H \left\langle \boldsymbol{\Sigma}^{-1} \right\rangle_{q_{\boldsymbol{\Sigma}}^{(j-1)}(\boldsymbol{\Sigma})} \mathbf{X}_i \right\}
\end{aligned} \tag{18}
$$

The exchange of soft-information between the data and parameter estimators is now complete with the complexity being similar to that of the equivalent EM-algorithm per iteration.

## 5. NUMERICAL EXAMPLE AND DISCUSSION

In order to indicate the advantage of the VBEM-algorithm and keep things as simple as possible, a single-antenna noise-limited GSM-like system is considered. The GSM system has a burst structure with $N_f = 142 + 6$ transmitted symbols, including the 6 so-called tailbits, and has $N_{tr} = 26$ known training symbols placed in the middle. The noise is assumed to be Additive White Gaussian Noise (AWGN) and the noise covariance estimation therefore reduces to a scalar variance estimation. The used channel model is the GSM Typical Urban (TU) multipath channel profile [9] with a speed of 0 km/h and using ideal frequency hopping. This ensures that the channel stays constant over a burst and that a new channel is drawn from the distribution for every burst, i.e. making it block-fading. The overall length of the transmission pulse-shaping and channel model is $L = 7$. To make a fair comparison with the EM-algorithm and not go into a discussion on the correctness of various choices of priors, only noninformative priors are used for the VBEM-algorithm, i.e. $\boldsymbol{\Sigma}_1^{-1} = \mathbf{0}$ and $\boldsymbol{\Sigma} \sim \mathcal{CW}^{-1}(0, \mathbf{0})$.

A difference between the considered system and a GSM system is, that the considered modulation is linearized in order to eliminate the non-linearities introduced by the GMSK modulation used in GSM. The resulting linear modulation is simply a BPSK modulation with a rotation of $\pi/2$ per symbol.

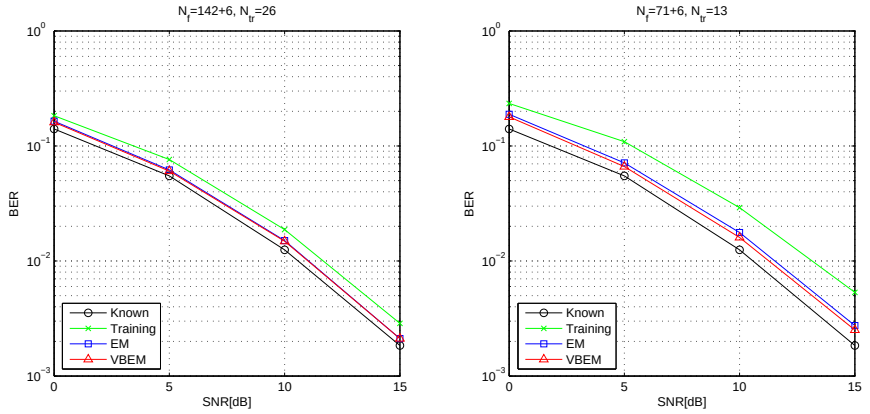


**Fig. 1**. Simulation of a GSM-like system using a TU0iFH channel profile, $N_t = N_r = 1$, $L = 7$.

On the left of Fig. 1, the Bit Error Rate (BER) of the above mentioned system is plotted. The results termed "Known" and "Training" are respectively the BER using the correct parameters and using only the training symbols for estimation. It can be seen that the BER of the EM and VBEM estimators are pretty much the same, although the VBEM estimator is actually better. The reason for this result is, that the number of observations is large compared to the number of parameters to be estimated. This makes the posterior distribution highly peaked around the ML solution effectively making the VBEM-algorithm fall back to the EM-algorithm.

However, changing the ratio between the number of estimated parameters and the number of observations affects the posterior distribution. On the right of Fig. 1, the length of the GSM burst has been reduced to half its original size leading to a less peaked posterior. The result is that the EM-algorithm now performs worse than the VBEM-algorithm, as the latter incorporates knowledge about the uncertainty in the parameters. The VBEM-algorithm is therefore beneficial when "few" observations are present or when "a lot" of parameters have to be estimated. This little example illustrates the advantage of the VBEM-algorithm for systems employing short packet structures and/or MIMO systems with many parameters to be estimated from a limited number of observations.

## 6. REFERENCES

[1] C. N. Georghiades and J. C. Han, "Sequence Estimation in the Presence of Random Parameters via the EM algorithm," *IEEE Trans. Commun.*, vol. 45, pp. 300–308, Mar. 1997.

[2] E. Chiavaccini and G. M. Vitetta, "MAP Symbol Estimation on Frequency-Flat Rayleigh Fading Channels Via a Bayesian EM Algorithm," *IEEE Trans. Commun.*, vol. 49, pp. 1869–1872, Nov. 2001.

[3] M. Nissil and S. Pasupathy, "Adaptive Bayesian and EM-Based Detectors for Frequency-Selective Fading Channels," *IEEE Trans. Commun.*, vol. 51, pp. 1325–1336, Aug. 2003.

[4] L. P. B. Christensen, "Minimum Symbol Error Rate Detection in Single-input Multiple-output Channels with Markov Noise," in *IEEE SPAWC Workshop*, 2005, pp. 236–240.

[5] N. M. Laird A. P. Dempster and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[6] M. Beal, *Variational Algorithms For Approximate Bayesian Inference*, Ph.D. thesis, 2003.

[7] M. Beal and Z. Ghahramani, "The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures," *Bayesian Stat., Oxford University Press*, vol. 7, pp. 453–464, 2003.

[8] C. P. Robert, *The Bayesian Choice*, Springer, 1994.

[9] 3GPP TS 45.005, *3GPP TSG GERAN; Radio transmission and reception (Release 5)*.

Appendix D

# An EM-algorithm for Band-Toeplitz Covariance Matrix Estimation

In Procedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1021-1024, 2007.

# AN EM-ALGORITHM FOR BAND-TOEPLITZ COVARIANCE MATRIX ESTIMATION

*Lars P. B. Christensen*

Technical University of Denmark
Informatics and Mathematical Modelling
email: lc@imm.dtu.dk

**ABSTRACT**

Toeplitz covariance matrix estimation has many uses in statistical signal processing due to the stationarity assumption of many signals. For some applications, further constraints may exist on the maximum lag at which the correlation function is non-zero and thereby giving rise to a band-Toeplitz covariance matrix. In this paper, an existing EM-algorithm for Toeplitz estimation is generalized to the case of band-Toeplitz estimation. In addition, the Cramer-Rao lower-bound for unbiased band-Toeplitz covariance matrix estimation is derived and through simulations it is shown that the proposed estimator achieves the bound for medium and large sample-sizes.

***Index Terms***— Structured covariance matrix estimation, banded Toeplitz, EM-algorithm, Cramer-Rao lower-bound.

## 1. INTRODUCTION

Estimation of Toeplitz covariance matrices is inherently connected to signal processing of stationary processes and applications are numerous, e.g. communications and radar systems. However, the constraint of stationarity and its resulting requirement for a Toeplitz structure in the covariance matrix, makes Maximum-Likelihood (ML) estimation challenging and no general closed-form solution is known [1, 2]. In [1], an EM-algorithm for Toeplitz covariance matrix estimation is constructed by exploiting a circulant extension of the Toeplitz matrix and the idea is further generalized to Block-Toeplitz in [3].

The contribution of this paper is to generalize the idea of using an EM-algorithm based on a circulant extension to covariance matrices that are not only Toeplitz, but band-Toeplitz with bandwidth $B$, i.e. having non-zero correlations only up to and including lag $B$. Such estimates are important in many practical applications as it is often reasonable to set an upper limit on the maximum lag of the estimate due to properties of the system considered. The proposed method therefore bridges the gap in correlation estimation from simple power estimation ($B = 0$) to full Toeplitz covariance matrix estimation. A natural generalization to block-banded block-Toeplitz matrices exists following [3], but this is outside the scope of this paper.

The author would like to thank Nokia Denmark for funding

Section 2 presents the system model and preliminaries and in Section 3 the EM-algorithm for Toeplitz covariance matrix estimation is outlined. Next, Section 4 modifies the existing M-step of the EM-algorithm to allow for constrained band-Toeplitz estimation. Section 5 derives the Cramer-Rao lower-bound for band-Toeplitz estimation and Section 6 outlines a traditional linear estimator used for comparison. Finally, Section 7 presents a numerical example and conclusion.

In the following, bold letters such as $\mathbf{x}$ and $\mathbf{X}$ are respectively column vectors and matrices with $x_i$ and $[\mathbf{X}]_{i,j}$ being a specific scalar element with the indices starting at zero. Further, $(\cdot)^T$ indicates matrix transpose, $(\cdot)^*$ matrix conjugation, $(\cdot)^H \triangleq \left((\cdot)^*\right)^T$ Hermitian transpose and $tr\{\cdot\}$ the trace operator. The notation $|\cdot|$ indicates the determinant of a matrix or the absolute value of a scalar. Finally, $diag(\cdot)$ constructs a diagonal matrix from a vector or, if operating on a matrix, produces a vector from the diagonal elements of the matrix.

## 2. SYSTEM MODEL AND PRELIMINARIES

Let $\mathcal{X} \triangleq \{\mathbf{x}_k\}_{k=1}^K$ be a collection of independent realizations of a zero-mean circular complex Gaussian distribution $\mathbf{x}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_x)$ with $\mathbf{x}_k \in \mathbb{C}^{N_x}$. The distribution of the observations is therefore given by

$$\begin{aligned} -ln\left(p\left(\mathcal{X} \mid \mathbf{\Sigma}_x\right)\right) - Kln\left|\pi\mathbf{\Sigma}_x\right| &= \sum_{k=1}^{K} \mathbf{x}_k^H \mathbf{\Sigma}_x^{-1} \mathbf{x}_k \\ &= Ktr\left\{\mathbf{\Sigma}_x^{-1}\mathbf{S}_x\right\} \end{aligned} \tag{1}$$

with $\mathbf{S}_x \in \mathbb{C}^{N_x \times N_x}$ being the sample covariance matrix

$$\mathbf{S}_x \triangleq \frac{1}{K}\sum_{k=1}^{K} \mathbf{x}_k \mathbf{x}_k^H \tag{2}$$

Also define $\mathbf{c} \in \mathbb{C}^{B+1}$ as the top $B+1$ terms of the first column in $\mathbf{\Sigma}_x$ and let $\mathbf{r} \in \mathbb{R}^{2B+1}$ be the stacking of the real and imaginary part of $\mathbf{c}$ with the real parts in the top of the vector. The length of $\mathbf{r}$ is only $2B+1$ as the imaginary part of $c_0$ must be zero.

Given the observations $\mathcal{X}$, the task is to provide the constrained ML estimate of $\mathbf{\Sigma}_x$. If no such constraints existed, the ML solution is simply given by $\mathbf{S}_x$.

Let $\hat{\mathbf{\Sigma}}_x^{(k)} \in \mathbb{C}^{N_x \times N_x}$ be an estimate of $\mathbf{\Sigma}_x$ at iteration $k$ that obeys the Toeplitz constraint. Now, let $\hat{\mathbf{\Sigma}}_y^{(k)} \in \mathbb{C}^{N_y \times N_y}$ be a circularly extended version of $\hat{\mathbf{\Sigma}}_x^{(k)}$ where $N_y \geq 2N_x - 1$ makes a circulant extension possible. It is well-known that the Discrete Fourier Transform (DFT) diagonalizes any circulant matrix, i.e.

$$\hat{\mathbf{\Sigma}}_y^{(k)} = \mathbf{D}^H \hat{\mathbf{\Lambda}}^{(k)} \mathbf{D} \tag{3}$$

with $[\mathbf{D}]_{i,j} = N_y^{-\frac{1}{2}} e^{-\frac{2\pi\sqrt{-1}}{N_y} ij}$ being the normalized DFT matrix of size $N_y$. The diagonal matrix $\hat{\mathbf{\Lambda}}^{(k)} \triangleq diag\left(\hat{\boldsymbol{\lambda}}^{(k)}\right)$ holds the eigenvalues given by

$$\hat{\boldsymbol{\lambda}}^{(k)} \triangleq N_y^{\frac{1}{2}} \mathbf{D} \hat{\boldsymbol{\sigma}}_y^{(k)} \tag{4}$$

where $\hat{\boldsymbol{\sigma}}_y^{(k)} \triangleq \left[\hat{\mathbf{\Sigma}}_y^{(k)}\right]_{:,0}$ is the first column of the circulant matrix. The eigenvalues of (4) contain all the information about the underlying Toeplitz matrix and we therefore have

$$\hat{\mathbf{\Sigma}}_x^{(k)} = \tilde{\mathbf{D}}^H \hat{\mathbf{\Lambda}}^{(k)} \tilde{\mathbf{D}} \tag{5}$$

with $\tilde{\mathbf{D}} \triangleq \mathbf{D}\left[\mathbf{I}_{N_x} \quad \mathbf{0}\right]^T$. If the desired covariance estimate is of size $M \leq N_x$, the result is given by the upper left submatrix of $\hat{\mathbf{\Sigma}}_x^{(k)}$.

## 3. EM-ALGORITHM FOR TOEPLITZ ESTIMATION

Here, the EM-algorithm applied to the problem of Toeplitz covariance matrix estimation is briefly outlined as described in [1] with [2] providing an efficient implementation. The E-step can be expressed as

$$E: \quad \mathbf{\Delta}^{(k)} \triangleq \hat{\mathbf{\Lambda}}^{(k)} \tilde{\mathbf{D}} \left(\mathbf{W}^{(k)} \mathbf{S}_x \mathbf{W}^{(k)} - \mathbf{W}^{(k)}\right) \tilde{\mathbf{D}}^H \hat{\mathbf{\Lambda}}^{(k)} \tag{6}$$

with $\mathbf{W}^{(k)} \triangleq \left(\hat{\mathbf{\Sigma}}_x^{(k)}\right)^{-1}$ and $\mathbf{\Delta}^{(k)}$ being the unconstrained update to the complete-data sample covariance matrix $\mathbf{S}_y$. The M-step should now choose the complete-data ML covariance estimate fulfilling the structural constraints based on the sufficient statistic $\mathbf{S}_y$. As the estimate is known to be Toeplitz, meaning that the update must be a diagonal matrix, it is straight-forward to show that the constrained ML update is exactly the diagonal of $\mathbf{\Delta}^{(k)}$, i.e.

$$M: \quad \hat{\mathbf{\Sigma}}_x^{(k+1)} = \tilde{\mathbf{D}}^H diag\left(\underbrace{\hat{\boldsymbol{\lambda}}^{(k)} + diag\left(\mathbf{\Delta}^{(k)}\right)}_{\hat{\boldsymbol{\lambda}}^{(k+1)}}\right) \tilde{\mathbf{D}} \tag{7}$$

As the EM-algorithm is only guaranteed to converge to a local maximum in the complete-data likelihood function, initialization is important. A reasonable choice of initialization, which is used throughout this paper, is $\hat{\mathbf{\Sigma}}_x^{(0)} = N_x^{-1} tr\{\mathbf{S}_x\} \mathbf{I}_{N_x}$.

## 4. A MODIFIED M-STEP FOR BAND-TOEPLITZ ESTIMATION

Assuming it is known apriori that the covariance matrix $\mathbf{\Sigma}_x$ is band-Toeplitz with bandwidth $B$, the idea is now to constrain the covariance estimate by requiring a functional form of $\hat{\boldsymbol{\lambda}}^{(k)}$ that guarantee this constraint. From (4) we have

$$\hat{\boldsymbol{\sigma}}_y^{(k)} = N_y^{-\frac{1}{2}} \mathbf{D}^H \hat{\boldsymbol{\lambda}}^{(k)} \tag{8}$$

so a set of eigenvalues $\hat{\boldsymbol{\lambda}}_{BT}^{(k)}$ fulfilling the structural constraint, must decompose as

$$\begin{aligned} \hat{\boldsymbol{\lambda}}_{BT}^{(k)} &= N_y^{\frac{1}{2}} \sum_{b=-B}^{B} \hat{c}_b^{(k)} \mathbf{d}_b \qquad s.t. \quad \hat{c}_b^{(k)} = \left(\hat{c}_{-b}^{(k)}\right)^* \\ &= N_y^{\frac{1}{2}} \left(\hat{c}_0^{(k)} \mathbf{d}_0 + 2 \sum_{b=1}^{B} Re\left\{\hat{c}_b^{(k)} \mathbf{d}_b\right\}\right) \end{aligned} \tag{9}$$

where $\mathbf{d}_b = \mathbf{d}_{-b}^* \triangleq [\mathbf{D}]_{:,mod(b,N_y)}$ with $mod(x,y)$ meaning $x$ modulo $y$. The functional form of the eigenvalues in (9) effectively forces the covariance estimate to be $\hat{\mathbf{c}}^{(k)} \triangleq \left[\hat{c}_0^{(k)}, \ldots, \hat{c}_B^{(k)}\right]^T$ for the non-zero band and zero elsewhere. However, as the eigenvalues must be real-valued and there are only $N_b \triangleq 2B+1$ real-valued degrees-of-freedom in the decomposition, we choose to reformulate the constraint as a real-valued decomposition, i.e.

$$\hat{\boldsymbol{\lambda}}_{BT}^{(k)} = \mathbf{T} \hat{\mathbf{r}}^{(k)} \tag{10}$$

Here $\hat{\mathbf{r}}^{(k)} \in \mathbb{R}^{N_b}$ are the unknowns and $\mathbf{T} \triangleq [\mathbf{t}_0, \ldots, \mathbf{t}_{N_b-1}]$ is defined by

$$\mathbf{t}_b \triangleq \begin{cases} N_y^{\frac{1}{2}} Re\{\mathbf{d}_b\} & , b = 0 \\ 2N_y^{\frac{1}{2}} Re\{\mathbf{d}_b\} & , 1 \leq b \leq B \\ -2N_y^{\frac{1}{2}} Im\{\mathbf{d}_{b-B}\} & , B < b < N_b \end{cases} \tag{11}$$

The upper $B+1$ coefficients of $\hat{\mathbf{r}}^{(k)}$ will therefore hold the real part of $\hat{\mathbf{c}}^{(k)}$ while the lower $B$ coefficients are the imaginary part. In the case of real-valued covariance estimation, having $N_b \triangleq B+1$ is therefore sufficient to parameterize the constrained estimate.

The reader should now be familiar with the overall structure of the EM-algorithm for band-Toeplitz estimation and the iteration index is therefore dropped for notational ease in the following. The challenge is now, given the current ML estimate over the space of circulant matrices $\hat{\boldsymbol{\lambda}}$, to minimize

some distance measure $f\left(\hat{\boldsymbol{\lambda}}, \hat{\mathbf{r}}\right)$ between the Toeplitz and the band-Toeplitz estimate subject to the constraint that the resulting eigenvalues must all be non-negative, i.e. $\mathbf{Tr} \geq \mathbf{0}$. As the EM-algorithm proceeds, this will result in a successive tightening of the lower-bound on the marginal log-likelihood determined in the E-step.

### 4.1. ML Estimation

Using the complete-data negative log-likelihood as a distance measure, we maximize the lower-bound over the space of valid band-Toeplitz matrices. Letting $\left[\tilde{\mathbf{t}}_0, \ldots, \tilde{\mathbf{t}}_{N_y-1}\right] \triangleq \mathbf{T}^T$, the desired distance measure can be written as

$$f\left(\hat{\boldsymbol{\lambda}}, \hat{\mathbf{r}}\right) = \sum_{i=0}^{N_y-1} ln\left(\tilde{\mathbf{t}}_i^T \hat{\mathbf{r}}\right) + \left(\tilde{\mathbf{t}}_i^T \hat{\mathbf{r}}\right)^{-1} \hat{\lambda}_i \tag{12}$$

As the distance measure consists of a sum of a concave and a convex term in the unknowns, the overall function is non-convex and thereby making global minimization unfeasible. Instead, the first- and second-order derivatives

$$\frac{\partial f\left(\hat{\boldsymbol{\lambda}}, \hat{\mathbf{r}}\right)}{\partial \hat{\mathbf{r}}} = \sum_{i=0}^{N_y-1} \frac{1}{\tilde{\mathbf{t}}_i^T \hat{\mathbf{r}}} \left(1 - \frac{\hat{\lambda}_i}{\tilde{\mathbf{t}}_i^T \hat{\mathbf{r}}}\right) \tilde{\mathbf{t}}_i \tag{13}$$

$$\frac{\partial^2 f\left(\hat{\boldsymbol{\lambda}}, \hat{\mathbf{r}}\right)}{\partial \hat{\mathbf{r}} \partial \hat{\mathbf{r}}^T} = -\sum_{i=0}^{N_y-1} \frac{1}{\left(\tilde{\mathbf{t}}_i^T \hat{\mathbf{r}}\right)^2} \left(1 - \frac{2\hat{\lambda}_i}{\tilde{\mathbf{t}}_i^T \hat{\mathbf{r}}}\right) \tilde{\mathbf{t}}_i \tilde{\mathbf{t}}_i^T \tag{14}$$

can be used in any favorite optimization scheme to determine a local minimum of the distance measure. As the distance measure may have multiple minima, the search should be started at the previous value of $\hat{\mathbf{r}}$ to make sure that the update cannot increase the distance measure.

### 4.2. Other Distance Measures

Instead of minimizing the negative log-likelihood function directly other criterions can also be used. However, for the EM-algorithm to converge an update must not increase (12), but it is not required to minimize it either. In this manner, it is possible to formulate an entire family of Generalized EM-algorithms for band-Toeplitz covariance estimation. An example of this strategy would be to not minimize (12), but only find an update that lowers it and thereby trade convergence speed for reduced computational complexity in the M-step.

## 5. LOWER-BOUND FOR BAND-TOEPLITZ COVARIANCE ESTIMATION

The Cramer-Rao Lower-Bound (CRLB) provides the lowest possible error variance of any estimator and is therefore a natural performance benchmark. Determining the bound for a biased estimator involves computing the bias-function of the estimator, which in general appears unfeasible. Instead, the Unbiased CRLB (U-CRLB) is derived and used for comparison, as it is well-known that the ML estimate is asymptotically unbiased.

Following the derivation in [4] and modifying it to include the complex-valued observations, the Fisher information matrix $\mathbf{J} \in \mathcal{R}^{N_b \times N_b}$ for the constrained covariance estimate can be found to be

$$[\mathbf{J}]_{i,j} = Ktr\left\{\boldsymbol{\Sigma}_x^{-1} \frac{\partial \boldsymbol{\Sigma}_x}{\partial r_i} \boldsymbol{\Sigma}_x^{-1} \frac{\partial \boldsymbol{\Sigma}_x}{\partial r_j}\right\} \tag{15}$$

and using (5) and (10) we readily get

$$\frac{\partial \boldsymbol{\Sigma}_x}{\partial r_i} = \tilde{\mathbf{D}}^H diag\left(\mathbf{t}_i\right) \tilde{\mathbf{D}} \tag{16}$$

As the focus is on the U-CRLB, the desired lower-bound is

$$E\left[\left|r_i - \hat{r}_i\right|^2\right] \geq \left[\mathbf{J}^{-1}\right]_{i,i} \tag{17}$$

## 6. WEIGHTED PROJECTED COVARIANCE ESTIMATION

This section outlines a simple method of performing Toeplitz covariance matrix estimation based on the idea of [5] in order to better understand the EM-based approach and provide a benchmark. The idea is to simply average along the diagonals of the sample covariance to estimate the correlations for the desired lags. However, to guarantee a positive definite matrix, the lag $m$ correlation estimate is weighted by $\frac{N_x - m}{N_x}$ with $m \geq 0$. A valid Toeplitz covariance matrix of size $M$ can now be constructed from the weighted correlation coefficients $\left[\hat{c}_0, \ldots, \hat{c}_{M-1}\right]^T$ resulting in a bias given by

$$E\left[c_m - \hat{c}_m\right] = \frac{m}{N_x} c_m \tag{18}$$

The Mean-Squared Error (MSE) of the correlation estimates can be found by expressing fourth-order moments of the Gaussian observations by their second-order moments as

$$E\left[\left|c_m - \hat{c}_m\right|^2\right] = \frac{m^2}{N_x^2}|c_m|^2 + \frac{1}{KN_x^2} \sum_{b=-B}^{B} \left(N_x - m - |b|\right) |c_{|b|}|^2 \tag{19}$$

The result of (19) is only valid for the complex-valued domain, but a similar result can be obtained for the real-valued domain by following the same principle. However, the proof of (19) and its real-valued equivalent are left out due to lack of space.

Although this sub-optimal method of covariance estimation can only provide full Toeplitz matrix estimates, it is related to band-Toeplitz estimation in the sense that one can choose $M = B + 1$ to produce an estimate of the non-zero correlation coefficients. Using the resulting estimate to produce a larger band-Toeplitz matrix, e.g. of size $N_x$, is however not guaranteed to be positive definite.

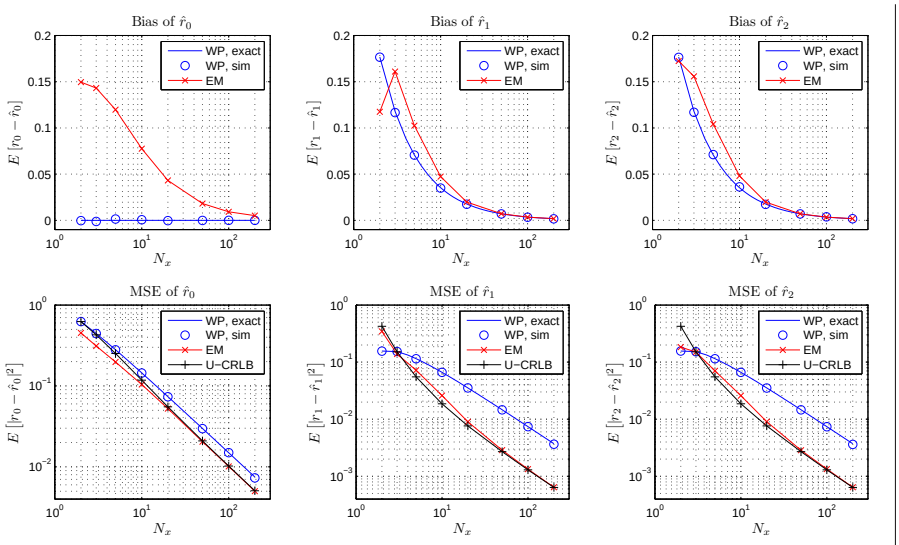


**Fig. 1**. Example of band-Toeplitz covariance matrix estimation as a function of the sample-size $N_x$ for $B = 1$, $K = 1$.

## 7. NUMERICAL EXAMPLE AND CONCLUSION

To demonstrate the proposed method, complex-valued zero-mean white Gaussian noise with unit power is filtered by a first-order FIR filter having coefficients $\left[\frac{1}{\sqrt{2}}, \frac{1+\sqrt{-1}}{2}\right]^T$. This results in a band-Toeplitz covariance matrix ($B = 1$) having the first column specified by $\mathbf{r} = \left[1, \frac{1}{\sqrt{8}}, \frac{1}{\sqrt{8}}\right]^T$. In figure 1, the bias and MSE of the Weighted Projected (WP) estimator and the proposed EM-based method (EM) is depicted as a function of the sample-size $N_x$. For WP, both exact results found by (19) and simulations are shown. The lower part of the figure depicts the MSE of the estimates and the U-CRLB is also shown for reference. It can be seen that a significant reduction in MSE is achievable by incurring a moderate bias increase with the exception of lag zero where WP is unbiased. Furthermore, as the sample-size increases, the proposed estimator becomes unbiased and tends to the U-CRLB. In conclusion, the proposed EM-based estimator provides near-optimal performance with a reasonable complexity due to its effective implementation exploiting the band-Toeplitz structure [2]. These properties make the proposed method an interesting candidate for many applications where accurate band-Toeplitz covariance matrix estimation is of great importance.

## 8. REFERENCES

[1] M. I. Miller and D. L. Snyder, "The role of likelihood and entropy in incomplete-data problems: Applications to estimating point-process intensities and toeplitz constrained covariances," *Procedings of the IEEE*, vol. 75, pp. 892–907, 1987.

[2] A. Dembo, C. L. Mallows, and L. A. Shepp, "Embedding nonnegative definite Toeplitz Matrices in nonnegative definite circulant matrices, with application to covariance estimation," *IEEE Trans. on Information Theory*, vol. 35, no. 6, pp. 1206–1212, Nov. 1989.

[3] D. R. Fuhrmann and T. A. Barton, "Estimation of block-toeplitz covariance matrices," in *24th IEEE Asim. Conf. on Signals, Systems and Computers.*, 1990, pp. 779–783.

[4] B. Porat and B. Friedlander, "Computation of the exact information matrix of Gaussian time series with stationary random components," *IEEE Trans. on Accoustics, Speech and Signal Proc.*, vol. 34, pp. 118–130, 1986.

[5] T. A. Barton and S. T. Smith, "Structured covariance estimation for space-time adaptive processing," in *Procedings of the IEEE ICASSP*, 1997, pp. 3493–3496.

# Bibliography


[3GP] TS 45.005 3GPP. *3GPP TSG GERAN; Radio transmission and reception (Release 5)*.

[ADC95] N. Al-Dhahir and J.M. Cioffi. MMSE decision-feedback equalizers: finite-length results. *IEEE Transactions on Information Theory*, 41(4):961–975, July 1995.

[AEVZ02] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger. Closest point search in lattices. *IEEE Transactions on Information Theory*, 48(8):2201–2214, Aug. 2002.

[AM00] S.M. Aji and R.J. McEliece. The generalized distributive law. *IEEE Trans. on Info. Theory*, 46(2):325–343, Mar. 2000.

[Att00] H. Attias. A variational bayesian framework for graphical models. In *Adv. Neur. Info. Proc. Sys. (NIPS) 12. MIT Press, Cambridge, MA.*, 2000.

[AV01] D. Agrawal and A. Vardy. The turbo decoding algorithm and its phase trajectories. *IEEE Transactions on Information Theory*, 47(2):699–722, Feb. 2001.

[BC02] J. Boutros and G. Caire. Iterative multiuser joint decoding: unified framework and asymptotic analysis. *IEEE Transactions on Information Theory*, 48(7):1772–1793, July 2002.

[BCJR74] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Transactions on Information Theory*, 20(2):284–287, Mar. 1974.

[Bea03] M. Beal. *Variational Algorithms For Approximate Bayesian Inference*. PhD thesis, 2003.

[BG03] M. Beal and Z. Ghahramani. The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. *Bayesian Stat., Oxford University Press*, 7:453–464, 2003.

[BGT93] C. Berrou, A. Glavieux, and P. Thitimajshima. Near Shannon limit error-correcting coding and decoding: Turbo-codes. In *IEEE Int. Conf. on Comm.*, volume 2, pages 1064–1070, May 1993.

[CC06] Michael Chertkov and Vladimir Y. Chernyak. Loop calculus helps to improve belief propagation and linear programming decodings of low-density-parity-check codes, 2006.

[Chr05a] L. P. B. Christensen. A Low-Complexity Joint Synchronization and Detection Algorithm for Single-Band DS-CDMA UWB Communications. *EURASIP, Journal on Applied Signal Processing, UWB - State of the Art*, 2005.

[Chr05b] L. P. B. Christensen. Minimum symbol error rate detection in single-input multiple-output channels with markov noise. In *IEEE SPAWC Workshop*, pages 236–240, June 2005.

[Chr07] L. P. B. Christensen. An EM-algorithm for Band-Toeplitz Covariance Matrix Estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007.

[CL06] L. P. B. Christensen and J. Larsen. On Data and Parameter Estimation using the Variational Bayesian EM-algorithm for Block-fading Frequency-selective MIMO Channels. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2006.

[DEGC03] M.O. Damen, H. El Gamal, and G. Caire. On maximum-likelihood detection and the search for the closest lattice point. *IEEE Transactions on Information Theory*, 49(10):2389–2402, Oct. 2003.

[DHH89] A. Duel-Hallen and C. Heegard. Delayed decision-feedback sequence estimation. *IEEE Transactions on Communications*, 37(5):428–436, May 1989.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[DMS89] A. Dembo, C. L. Mallows, and L. A. Shepp. Embedding nonnegative definite Toeplitz Matrices in nonnegative definite circulant matrices, with application to covariance estimation. *IEEE Transactions on Information Theory*, 35(6):1206–1212, Nov. 1989.

[EKM06] M. El-Khamy and R.J. McEliece. Iterative algebraic soft-decision list decoding of Reed-Solomon codes. *IEEE Journal on Selected Areas in Communications*, 24(3):481–490, Mar. 2006.

[EQ88] M.V. Eyuboglu and S.U.H. Qureshi. Reduced-state sequence estimation with set partitioning and decision feedback. *IEEE Transactions on Communications*, 36(1):13–20, Jan. 1988.

[FB90] D. R. Fuhrmann and T. A. Barton. Estimation of block-toeplitz covariance matrices. In *24th IEEE Asilomar Conference on Signals, Systems and Computers.*, pages 779–783, Nov. 1990.

[FN02] T. Fabricius and O. Nørklit. Approximations to joint-ml and ml symbol-channel estimators in mud cdma. In *IEEE Global Telecommunications Conference (GLOBECOM)*, pages 389–393, Nov. 2002.

[GG84] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[GH03] R. Gowaikar and B. Hassibi. Efficient statistical pruning for maximum likelihood decoding. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 49–52, April 2003.

[GOS+04] W.H. Gerstacker, F. Obernosterer, R. Schober, A.T. Lehmann, A. Lampe, and P. Gunreben. Equalization Concepts for Alamouti's Space-Time Block Code. *IEEE Transactions on Communications*, 52(7):1178–1190, July 2004.

[GSL03] W.H. Gerstacker, R. Schober, and A. Lampe. Receivers with widely linear processing for frequency-selective channels. *IEEE Transactions on Communications*, 51(9):1512–1523, Sep. 2003.

[Hay96] S. Haykin. *Adaptive Filter Theory.* Prentice-Hall, 3rd edition, 1996.

[HtB03] B.M. Hochwald and S. ten Brink. Achieving near-capacity on a multiple-antenna channel. *IEEE Transactions on Communications*, 51(3):389–399, Mar. 2003.

[HV05a] B. Hassibi and H. Vikalo. On the sphere-decoding algorithm I. Expected complexity. *IEEE Transactions on Signal Processing*, 53(8):2806–2818, Aug. 2005.

[HV05b] B. Hassibi and H. Vikalo. On the sphere-decoding algorithm II. Generalizations, second-order statistics, and applications to communications. *IEEE Transactions on Signal Processing*, 53(8):2819–2834, Aug. 2005.

[KFL01] F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. on Info. Theory*, 47(2):498–519, Feb. 2001.

[KFSW02] C. Komninakis, C. Fragouli, A. H. Sayed, and R. D. Wesel. Multi-input multi-output fading channel tracking and equalization using Kalman estimation. *IEEE Trans. on Signal Processing*, 50(5):1065–1076, May 2002.

[KM00] A. Kavcic and J.M.F. Moura. Matrices with banded inverses: inversion algorithms and factorization of Gauss-Markov processes. *IEEE Transactions on Information Theory*, 46(4):1495–1509, July 2000.

[KST04] R. Koetter, A.C. Singer, and M. Tuchler. Turbo equalization. *IEEE Signal Processing Magazine*, 21(1):67–80, Jan. 2004.

[LR04] T. Lin and L.K. Rasmussen. Application of maximal ratio combining for iterative multiuser decoding. In *International Symposium on Information Theory*, page 531, June 2004.

[LR05] T. Lin and L.K. Rasmussen. Truncated maximal ratio combining for iterative multiuser decoding. In *Proceedings of the 6th Australian Communications Theory Workshop*, pages 35–41, Feb. 2005.

[Mac97] D. J. C. MacKay. Ensemble learning for hidden Markov models. `www.inference.phy.cam.ac.uk/mackay/abstracts/ensemblePaper.html`, 1997.

[MMC98] R.J. McEliece, D.J.C. MacKay, and Jung-Fu Cheng. Turbo decoding as an instance of Pearl's "belief propagation" algorithm. *IEEE Journal on Sel. Areas in Comm.*, 16(2):140–152, Feb. 1998.

[MN97] D.J.C. MacKay and R.M. Neal. Near Shannon limit performance of low density parity check codes. *Electronics Letters*, 33(6):457–458, Mar. 1997.

[MS87] M. I. Miller and D. L. Snyder. The role of likelihood and entropy in incomplete-data problems: Applications to estimating point-process intensities and toeplitz constrained covariances. *Proceedings of the IEEE*, 75:892–907, July 1987.

[NH99] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, pages 355–368, 1999.

[NP03] M. Nissilä and S. Pasupathy. Adaptive Bayesian and EM-based detectors for frequency-selective fading channels. *IEEE Trans. on Commun.*, 51(8):1325–1336, Aug. 2003.

[OT04] R. Otnes and M. Tuchler. Iterative channel estimation for turbo equalization of time-varying frequency-selective channels. *IEEE Transactions on Wireless Communications*, 3(6):1918–1923, Nov. 2004.

[PA06] P. Pakzad and V. Anantharam. Kikuchi Approximation Method for Joint Decoding of LDPC Codes and Partial-Response Channels. *IEEE Transactions on Communications*, 54(7):1149–1153, July 2006.

[Poo88] H. V. Poor. *An introduction to signal detection and estimation*. Springer, 1988.

[Pro95] J. G. Proakis. *Digital Communications*. McGraw-Hill, 3rd edition, 1995.

[PV97] H.V. Poor and S. Verdu. Probability of error in MMSE multiuser detection. *IEEE Transactions on Information Theory*, 43(3):858–871, May 1997.

[Sha48] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, July 1948.

[SK01] A. H. Sayed and T. Kailath. A survey of spectral factorization methods. *Numerical linear algebra with applications*, 8(6-7):467–496, 2001.

[SWS04] O. Shental, A. J. Weiss, and Y. Shental, N. Weiss. Generalized belief propagation receiver for near-optimal detection of two-dimensional channels with memory. In *Information Theory Workshop*, pages 225–229, October 2004.

[Tel99] I. E. Telatar. Capacity of multi-antenna Gaussian channels. *European Transactions on Telecommunications*, 10(6):585–595, Nov./Dec. 1999. Also in AT&T Bell Lab. Tech. Memo, June 1995.

[TM03] M. Tüchler and M. Mecking. App equalization for non-ideal channel knowledge. In *IEEE International Symposium on Information Theory*, pages 356,, 2003.

[TSK02] M. Tuchler, A.C. Singer, and R. Koetter. Minimum mean squared error equalization using a priori information. *IEEE Transactions on Signal Processing*, 50(3):673–683, Mar. 2002.

[TV05] D. Tse and P. Viswanath. *Fundametals of Wireless Communication*. Cambridge University Press, 2005.

[VHK04] H. Vikalo, B. Hassibi, and T. Kailath. Iterative decoding for MIMO channels via modified sphere decoding. *IEEE Transactions on Wireless Communications*, 3(6):2299–2311, Nov. 2004.

[VHM06] H. Vikalo, B. Hassibi, and U. Mitra. Sphere-constrained ML detection for frequency-selective channels. *IEEE Transactions on Communications*, 54(7):1179–1183, July 2006.

[Vit67] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, Apr. 1967.

[WMT05] M. Welling, T. Minka, and Y. W. Teh. Structured region graphs: Morphing EP into GBP. In *21st Conf. on Uncertainty in Artificial Intelligence (UAI)*, Juli 2005.

[XES01] G. Xianping, D. Eppstein, and P. Smyth. The distribution of loop lengths in graphical models for turbo decoding. *IEEE Transactions on Information Theory*, 47(6):2549–2533, Sep. 2001.

[XZ04] C. Xiao and Y.R. Zheng. Ergodic capacity of mimo triply selective rayleigh fading channels. In *IEEE Global Telecommunications Conference (GLOBECOM)*, volume 5, pages 3133–3137, Nov. 2004.

[YFW05] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. on Info. Theory*, 51(7):2282–2311, July 2005.

[Yui02] A. L. Yuille. Cccp algorithms to minimize the bethe and kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, 14(7):1691–1722, 2002.

[ZT03] L. Zheng and D.N.C. Tse. Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels. *IEEE Transactions on Information Theory*, 49(5):1073–1096, May 2003.